Identification and Estimation of a Triangular model with Multiple Endogenous Variables and Insufficiently Many Instrumental Variables*

Liquan Huang† Umair Khalil‡ Neşe Yıldız§

First version: April 2013. This version: July 31, 2015.

Abstract

We develop a novel identification method for a partially linear model with multiple endogenous variables of interest in which the number of available instruments is strictly less than the number of endogenous variables. We present an easy-to-implement consistent estimator for the parametric part. We show that this estimator retains $\sqrt{n}$-convergence rate and asymptotic normality even with the presence of generated regressors. The nonparametric part of the model is also identified. Monte Carlo simulation demonstrates that our estimator performs well in finite samples. We use our methods to assess the impact of smoking during pregnancy on birth weight.

JEL Classifications: *C31, C36, C13, C14*

Key words: Identification, partially linear model, control function approach.

*We are grateful to Bin Chen, Carol Caetano, Greg Caetano and conference participants at New York Camp Econometrics VIII for their helpful comments and discussions. Any remaining errors are ours.

†Department of Economics, University of Rochester, 231 Harkness Hall, Rochester, NY 14627; Email: l.huang@rochester.edu.

‡Department of Economics, University of Rochester, 231 Harkness Hall, Rochester, NY 14627; Email:ukhalil@mail.rochester.edu

§Corresponding author: Department of Economics, University of Rochester, 231 Harkness Hall, Rochester, NY 14627; Email: nese.yildiz@rochester.edu; Phone: 585-275-5782; Fax: 585-256-2309.

1 Introduction

This paper considers models with multiple sources of endogeneity in which the number of potential instrumental variables is strictly less than the number of endogenous regressors of interest. We show how parameters of interest in such models can be identified. We also provide a new estimator that is easy to implement for the coefficient of the endogenous regressor for which no instrumental variables are available. By applying the empirical process methods, we show that the estimator retains $\sqrt{n}$-convergence rate and is asymptotically normal even with the presence of generated regressors. The nonparametric part of the model is also identified, and can be estimated with the standard nonparametric convergence rate.

To see why methods developed in this paper might be interesting, consider the problem of studying the dynamic evolution of crime. That is, suppose we are interested in estimating how crime in last period affects crime in current period. Jacob, Lefgren and Moretti (2007) use a variable measuring last period's weather conditions as an IV for last period's criminal activity. Using this instrument they estimate the effect of last period's crime on this period's crime. But this effect might be a mixture of different effects. For example, part of the effect of last period's crime on this period's crime might be due to learning-by-doing. Part of the effect might be the result of other channels. For instance, police response in a given neighborhood, might also respond to last period's criminal activity making inference difficult. Moreover, we might also be interested in evaluating the causal effect of last period's criminal activity through such additional channels themselves. This, in turn, might be challenging as these channels might also be endogenous. If police response is also related to weather conditions and if current weather is related to last period's weather then weather conditions may not be a valid IV. Even if we believe the validity of this IV, we cannot use the IV method to estimate the effect of last period's crime on this period's crime through its effect on police response, given the latter might be correlated with unobservables. The methods developed in this paper can be used to evaluate exactly such effects.

In particular, suppose Y denotes a measure of this period's criminal activity, X denotes last period's criminal activity, Z denotes a measure of last period's weather, D denotes other endogenous variables, like a measure of policing activity (which could be continuous or discrete) that we might be interested in, and ε represents the remaining unobservables. The model we study is

$$Y = \lambda(X) + D^{\intercal}\gamma + \varepsilon, \tag{1}$$

$$X = \pi(Z) + V. \tag{2}$$

By running a first stage non-parametric regression of the endogenous regressor X on the instrument we could identify V. The crucial assumption we impose is that the residual from this regression will be a control function for the endogenous variable for which an IV is available. That is, we assume $\mathbb{E}(\varepsilon|X,V) = \mathbb{E}(\varepsilon|V) =: \rho(V)$. Note that inclusion of V into the outcome equation as a regressor will only control for endogeneity of X, but not of D. This means that we could write $\varepsilon = \rho(V) + \varsigma$, where $\mathbb{E}(\varsigma|X,V) = 0$ by construction. In particular, X and V will be additively separable and exogenous in the outcome equation written with ς as the only unobserved term. Note, however, that $E(\varsigma|D)$ is not required to be 0.Then the cross partial derivative of the conditional expectation of the outcome given X, V with respect to X and V must be equal to γ (the coefficient vector on the endogenous regressors D) times the cross partial derivative of the conditional expectation of D given X, V with respect to X and V. If the vector of cross partial derivatives of $\mathbb{E}(D|X,V)$ with respect to X and V are linearly independent, which is a testable assumption, γ, the coefficient of D, is identified. Once γ is identified, we can subtract $D^\intercal\gamma$ from the outcome and identify $\lambda(X)$ as well as $\rho(V)$ up to some location normalization. Note that if the instrument Z is binary we could only identify the continuous unobservable (but identifiable) control function V at two points. In that case, the coefficient vector γ can still be identified by replacing partial derivatives with differencing, so that cross partial derivatives are replaced by differences in differences. Full identification of $\lambda(X)$ and $\rho(V)$, however, will not really be feasible if the instrument is binary. With a single binary instrument, only two values, v_1 and v_0, of V will be identified. We can still identify $\frac{\partial\lambda(x)}{\partial x}$ for x in the support of $X|V = v_1$ or $X|V = v_0$ if the instrument Z is binary.

The tools we develop may be useful in many other estimation problems as well. For example, consider estimation of the effect of maternal smoking during pregnancy on birth weight. Medical literature has long established a link between maternal smoking and adverse birth outcomes, primarily measured by the effect on average birth weights. However, the primary concerns involved in estimating the causal effect of maternal smoking on birth weights is the lack of an exogenous source of variation. The most comprehensive study on the question, Almond et al. (2005) also use a selection-on-observables framework. It is particularly hard to find a source of variation that is correlated with smoking during pregnancy but uncorrelated with unobserved 'mothering ability'. Using the methods proposed in this paper, we can potentially solve this issue by using an available IV for a second endogenous variable of interest in our structural equation. For instance, consider the investigation by Currie and Moretti (2003) of the effect of maternal education on birth outcomes. Given the former is clearly endogenous, they use the number of colleges in the county where the mother was resident at age 17 as their instrumental variable. Under

this setup, we estimate the causal effect of smoking on birth weights in Section 6. Hence, our method can be useful in such instances as well where the lack of an IV for an endogenous variable of interest can be overcome if an IV is available for a second endogenous regressor in the structural equation.

Based on our identification strategy, we propose an easy-to-compute estimator which achieves $\sqrt{n}$-convergence rate for γ. Asymptotic normality of this estimator is also derived. We should point out that our approach consists of a multi-step estimation and applies the control function approach which results in having a generated regressor. The consistency and $\sqrt{n}$-convergence result is non-trivial when generated regressors are involved.

Partially linear model has been well established in econometric theory since the seminal work of Robinson (1988) and Speckman (1988). The identification and estimation of partially linear model with endogeneity in either the parametric part or the nonparametric part have been discussed in the literature (e.g., Ai and Chen 2003, Chen and Pouzo 2009, Florens 2003). A recent study, Florens et al. (2012) also propose a $\sqrt{n}$-consistent estimator for the parametric part, when endogeneity exists in both the parametric and nonparametric parts. However their method relies on the availability and strength of the IV. Our method only requires the availability of IV for the nonparametric part, and the endogenous variable in the linear part of our model is allowed to be binary or censored.

Triangular equations model is a common way to model endogeneity especially in treatment effect literature. To name a few, Imbens and Newey (2009) and Newey et al. (1999) both focus on the identification and estimation of a two equation triangular model. Both methods require the existence of sufficient IVs. Our paper also considers a triangular model, but requires fewer IVs and allows both continuous and discrete variables in the linear part of the model.

This paper is also related to the control function approach (see Blundell and Powell, 2003). In our case, however, the control function is not required to control for all sources of endogeneity; its inclusion into the outcome equation is only required to control for endogeneity in X, but not in D. We use a first stage nonparametric regression of X on Z to obtain this control function. But our methods could easily be extended to the case where X is a strictly monotone function of V, the first stage structural unobservable, as in Imbens and Newey (2009).

Based on our identification strategy, the proposed estimator is of the ratio form, where both the numerator and the denominator are averages of cross partial derivatives of nonparametric regressions with generated regressors. The generated regressor literature is growing rapidly in recent years, including Mammen et al. (2012a, 2012b), Hahn and Ridder (2013), and Escanciano et al. (2014). The way

we handle the generated regressor is mostly related to Mammen et al. (2012a). They use the local linear regression to simplify technical arguments and focus on the conditional mean with the generated regressor. We consider the averages of cross partial derivatives of nonparametric regressions, and apply local polynomial regression to obtain estimates of the cross-partial derivatives. Lee (2014) also studies the averages of such regression by using the kernel estimation. However, we have a different focus, and our main contribution is $\sqrt{n}$-consistency and asymptotic normality result for the parametric estimator in an under-identified partially linear model. A by-product gained in our estimation is that, as an intermediate step, the estimator of the average of the second order derivatives is shown to be $\sqrt{n}$-consistent. Li et al. (2003) show that the estimator of the average of first order derivatives converges at parametric rate. Our paper extends their results to the second order derivatives. Instead of using U-statistics, we apply the empirical process methods to address the problem.

The paper is organized as follows. In Section 2, we introduce the model and the identification strategy. We also demonstrate our crucial identification assumption is not restrictive and list a few examples. Section 3 proposes the estimators and focuses on the asymptotic property of the parametric part. Section 4 discusses a few possible extensions of the basic model. These extensions might be more relevant in certain empirical applications. In Section 5, the Monte Carlo simulation study indicates our estimators perform well in finite samples. Section 6 provides an empirical study which illustrates how our method can be applied. We conclude in Section 7. The proof of the main results is deferred to the Mathematical Appendix.

2 The Model and the Identification

Consider a partially linear model

$$Y = \lambda(X_1, Z_1) + \gamma D + \varepsilon\,, \tag{3}$$

where Y is a scalar and observable, X_1 is a vector of observed variables in $\mathcal{R}^{d_x}$, Z_1 is a vector of exogenous variables in $\mathcal{R}^{d_1}$, D is a univariate observed variable, and ε is an unobserved variable. ε is possibly correlated with D. In our model, we assume D is endogenous and can be continuous, binary, censored and so on. For simplicity, we consider only univariate D. However, our approach can easily be extended to multivariate case.

Our model covers two cases: (i) endogeneity in parametric part only, i.e. $k = 1$; (ii) endogeneity in both parametric and nonparametric parts, i.e. $k > 1$, although the second case is the more interesting

case. For case (i), X_1 is assumed to be exogenous, which leaves D the only endogenous variable in the model. For case (ii), X_1 is also endogenous. For case (ii) additional notation need to be introduced, as we assume there exist IVs for X_1. To avoid unnecessary confusion, we first focus on case (ii), and revisit case (i) later in Section 2.1,

For case (ii), X_1 is supposed to be endogenous, and we assume there exists a vector of excluded instrumental variables Z_2 in $\mathcal{R}^{d_2}$. Our model becomes a triangular semiparametric model:

$$\begin{aligned} Y &= \lambda(X_1, Z_1) + \gamma D + \varepsilon\,, \\ X_1 &= \pi(Z_1, Z_2) + V\,. \end{aligned} \tag{4}$$

Here, V is a $d_x \times 1$ residual from the nonparametric population regression, which serves as the control variable for X_1 and we have $\mathbb{E}(V|Z_1, Z_2) = 0$. Let $X = (X_1, Z_1)$, $Z = (Z_1, Z_2)$, $d_z = d_1 + d_2$ and $d = d_1 + d_x$. Now we introduce one of our main identifying assumptions.

Assumption 1 *Suppose* $\mathbb{E}(\varepsilon|X, V) = \mathbb{E}(\varepsilon|V)$.

Assumption 1 is commonly seen in control function approach literature and also imposed in Newey et al. (1999). This assumption states that V is a control function for X. Under Assumption 1, we could write

$$\varepsilon = \rho(V) + \varsigma\,, \tag{5}$$

where $\rho(V) = \mathbb{E}(\varepsilon|V)$. Then, $\mathbb{E}(\varsigma|X, V) = 0$ by construction.

To avoid notational complexity, we assume $d_x = 1$. We also assume $d_2 = 1$, so we only have one instrumental variable for the endogenous variable X. The analysis with $d_x > 1$ and $d_2 > 1$ is similar. We impose the following regularity conditions,

Assumption 2 $(Y_i, X_{1i}, Z_{1i}^\intercal, Z_{2i})$ *are independent and identically (i.i.d.) distributed as* $(Y, X_1, Z_1^\intercal, Z_2)$. X *is a* $d \times 1$ *random vector and* Z *is a* $d_z \times 1$ *random vector. Suppose* (X, Z_2) *has compact and convex support* $\mathcal{X} \times \mathcal{Z}_2 \subset \mathcal{R}^{d+d_2}$.

Assumption 3 *(i)* $\mathbb{E}(Y|X, V)$ *and* $\mathbb{E}(D|X, V)$ *have continuous derivatives of total order* $p + 1$. *Let* $f(x, v)$ *denote the density function of* (X, V). $f(\cdot)$ *also has continuous derivatives of total order* 3, *and* $\inf_{(x,v)\in\mathcal{X}\times\mathcal{V}} f(x, v) \geq \eta$ *for some* $\eta > 0$*; (ii)* $\pi(Z)$ *is second order continuously differentiable.*

Assumption 4 $\mathbb{E}\left[\frac{\partial^2 \mathbb{E}(D|X,V)}{\partial X_1 \partial V}\right] \neq 0$.

Assumption 2 is very standard in the literature, and the i.i.d assumption can be relaxed to weakly dependence which requires more complicated and tedious derivation. Assumption 3 assumes some smoothness condition on both the conditional expectation and the joint density function of (X, V), which is widely used in local polynomial regression literature. Assumption 4 is the crucial assumption in our identification strategy. This assumption is quite general and allows D to be continuous, binary and censored. All it requires is that the reduced-form regression of D on X and V cannot be purely additive in X_1 and V, but also contains a cross term. This reflects the existence of an interaction effect of X_1 and V only through the endogenous variable D, which serves as the additional exogenous variation we need.

Let $u = D - \mathbb{E}(D|X, V)$. Under Assumption 1, we can write the model as a three-equation model,

$$\begin{aligned} Y &= \lambda(X) + \gamma D + \rho(V) + \varsigma\,, \\ D &= \mathbb{E}(D|X,V) + u\,, \\ X_1 &= \pi(Z) + V\,. \end{aligned} \tag{6}$$

After we control for V, the endogeneity problem caused by X_1 no longer exists. However, the model is still subject to endogeneity issue, as D can be correlated with ς. Let $\xi = \varsigma + \gamma u$. Our model becomes

$$\begin{aligned} Y &= \lambda(X) + \gamma \mathbb{E}(D|X,V) + \rho(V) + \xi\,, \\ \mathbb{E}(Y|X,V) &= \lambda(X) + \gamma \mathbb{E}(D|X,V) + \rho(V)\,. \end{aligned} \tag{7}$$

By the difference of the first equation in (6) and (7), we have

$$Y - \mathbb{E}(Y|X,V) = \gamma[D - \mathbb{E}(D|X,V)] + \varsigma\,,$$

and D is still endogenous. Thus, Robinson's (1988) method doesn't apply here if no instrumental variables for D exist. Therefore, we propose a new identification strategy. By Assumption 3 and equation (7), we have

$$\frac{\partial^2 \mathbb{E}(Y|X,V)}{\partial X_1 \partial V} = \gamma \frac{\partial^2 \mathbb{E}(D|X,V)}{\partial X_1 \partial V}\,.$$

Actually, we only require $\mathbb{P}\left(\frac{\partial^2 \mathbb{E}(D|X,V)}{\partial X_1 \partial V} \neq 0\right) > 0$ to have γ identified. For the purpose of estimation, a

relatively stronger condition is given by Assumption 4. Then, it's straight-forward to see that

$$\gamma = \mathbb{E}\left[\frac{\partial^2 \mathbb{E}(Y|X,V)}{\partial X_1 \partial V}\right] \Big/ \mathbb{E}\left[\frac{\partial^2 \mathbb{E}(D|X,V)}{\partial X_1 \partial V}\right], \tag{8}$$

so γ is identified.

When γ is identified, by (7), we have

$$\mathbb{E}(Y - \gamma D|X,V) = \lambda(X) + \rho(V) =: \tilde{\lambda}(X,V)\,, \tag{9}$$

then $\tilde{\lambda}(X,V)$ is identified. To identify $\lambda(\cdot)$ and $\rho(\cdot)$ up to a constant, we may use

$$\frac{\partial \mathbb{E}(Y - \gamma D|X = x, V = v)}{\partial x} = \frac{d\lambda(x)}{dx}$$

and

$$\frac{\partial \mathbb{E}(Y - \gamma D|X = x, V = v)}{\partial v} = \frac{d\rho(v)}{dv}\,.$$

The identification and estimation of the nonparametric part are quite standard and not the focus of our paper.

Assumption 4 is a rank condition like the relevance condition in the standard IV method. The rank condition in the standard IV model says that the exogenous variations in the instrument should lead to variations in the endogenous variable in the model, which also lead to variations in outcome. Here, we do not assume that there is a valid IV for D. Given the structure of the model, however, when we change X and then V the resulting change in conditional average outcome given X and V must be due to the change in conditional average D given X and V. Thus, we can only identify the coefficient on D if the change in conditional average D given X and V when both X and V change in order is not zero. Below, we provide two common situations where this change could be expected to be not zero.

Example 2.1 - Continuous D: Consider D generated by a continuous variable X and a disturbance ε_2,

$$D = X'\beta + \varepsilon_2\,.$$

Then $\mathbb{E}(D|X,V) = X'\beta + \mathbb{E}(\varepsilon_2|X,V)$. As long as $\mathbb{E}(\varepsilon_2|X,V) \neq \mathbb{E}(\varepsilon_2|V)$, we can identify γ. In this case, our identification strategy requires V to be a valid control for X only but not for D.

Moreover, Assumption 4 can be easily satisfied especially when D is binary or censored.

Example 2.2 - Binary D:

$$D = 1\{X_1^\intercal \theta_1 + Z_1^\intercal \theta_2 \geq \eta\}\,.$$

Then $\mathbb{E}(D|X, V, Z_1) = F_{\eta|X,V,Z_1}(x_1^\intercal\theta_1 + z_1^\intercal\theta_2|x, v, z_1)$, where η represents the unobservables, and $F_{\eta|X,V,Z_1}$ denotes the conditional distribution of this variable given X, V and Z_1. Using the strategy we propose, we could identify all the parameters in the outcome equation even if $\eta|X, Z_1, V \sim \eta|V$. If, however, this additional condition holds then the methods we provide here can be combined with the methods provided in Blundell and Powell (2004), to identify all the parameters of the model.

2.1 Case (i): $k = 1$

When only D is endogenous, we are able to identify its marginal effect on Y without the availability of any IVs. To identify the coefficient on D we need at least two exogenous regressors which have an interaction effect only through D. While we think identification of γ in this case is a nice theoretical result, we expect that our results with endogenous X to be more relevant in empirical applications. To avoid any confusion caused by the changes of notation, we restate our model when $k = 1$. Our model now is a single-equation model,

$$Y = \lambda(X) + \gamma D + \rho(V) + \varsigma\,,$$

where Y is still a scalar and observable, D is an endogenous observed variable, and $(X^\intercal, V)$ is a vector of exogenous and observed variables. The crucial identification assumption is the same as Assumption 4, and the intuition stays the same. It is straightforward to see that under Assumption 4, equations (7) and (8) still hold, and γ is identified. The nonparametric part can also be identified similarly as in case (ii). The earlier listed examples also carry over for this case. When the only endogenous variable D is continuous, our identification strategy can be applied to any model that falls into the category of the example below.

Example 2.3:

$$Y = \lambda(X) + \lambda D + \rho(V) + \varsigma$$
$$\mathbb{E}(D|X = x, V = v) = h(x, v)\,,$$

where $h(x, v)$ cannot be additive separable in X and V.

3 Estimation

We propose an estimator for γ based on (8) for both cases $k > 1$ and $k = 1$. When only D is endogenous and there is no IV, i.e. $k = 1$, the estimation procedure is much simpler. When both X and D are endogenous, and we only have IV for X ($k > 1$), the estimation is more involved due to the generated regressor $\hat{V}$. Let's start with the more complicated case that both X and D are endogenous but only Z, the IV for X, is available. We shall revisit case (i) later. The estimation procedure for case (ii) consists of three stages,

1. We run local linear regression of X_1 on Z for the first stage regression (4) to get $\hat{\pi}(\cdot)$. Subtracting $\hat{\pi}(Z)$ from X_1, we obtain the residual $\hat{V}$,

$$\hat{V} = X_1 - \hat{\pi}(Z)\,,$$

   where $\hat{\pi}(Z) = \hat{\mathbb{E}}(X|Z) = \hat{\alpha}_X$ is defined later by (11).

2. We use local polynomial regression to obtain the estimators $\frac{\widehat{\partial^2 \mathbb{E}}(Y|X,\hat{V})}{\partial X_1 \partial V}$ and $\frac{\widehat{\partial^2 \mathbb{E}}(D|X,\hat{V})}{\partial X_1 \partial V}$, which are later defined by (12) and (13). These estimate the partial derivatives $\frac{\partial^2 \mathbb{E}(Y|X,V)}{\partial X_1 \partial V}$ and $\frac{\partial^2 \mathbb{E}(D|X,V)}{\partial X_1 \partial V}$ with the generated regressor $\hat{V}$.

3. Finally, we calculate the sample average for the unconditional moment using the estimators from step two,

$$\hat{\gamma} = \hat{\mathbb{E}}\left[\frac{\widehat{\partial^2 \mathbb{E}}(Y|X,\hat{V})}{\partial X_1 \partial V}\right] \Big/ \hat{\mathbb{E}}\left[\frac{\widehat{\partial^2 \mathbb{E}}(D|X,\hat{V})}{\partial X_1 \partial V}\right], \tag{10}$$

   where $\hat{\mathbb{E}}[\cdot]$ is the sample average.

As we can see, the major challenge here is to solve how the generated regressor influences our final

estimator. Before we formally define the local linear estimator and local polynomial estimators listed above, we first introduce the assumption imposed.

Assumption 5 *The kernel function $K(\cdot)$ is a non-negative, twice continuously differentiable function with a compact support and satisfies $\int K(u)\,du = 1$ and $\int uK(u)\,du = 0$. For any multivariate vector $\nu \in \mathcal{R}^{d_\nu}$, define $K(\nu) = K(\nu_1)K(\nu_2)\cdots K(\nu_{d_\nu})$.*

Assumption 5 is very standard in nonparametric estimation. Here we state the assumption for any kernel used in the remainder of the paper. We use product kernels whenever there are multiple (continuous) conditioning variables. There are several widely used kernel functions satisfying Assumption 5, such as bi-weight kernel.

Here and in the sequel, denote e_t as a $\mathbf{N}$-dimensional[1] unit vector with 1 at the t^{th} argument. Let $\frac{\widehat{\partial^2\mathbb{E}}(Y|X,\hat{V})}{\partial X_1 \partial V} = e_{2d+3}^{\intercal}\hat{\beta}_Y$ and $\frac{\widehat{\partial^2\mathbb{E}}(D|X,\hat{V})}{\partial X_1 \partial V} = e_{2d+3}^{\intercal}\hat{\beta}_D$. We have $\hat{\alpha}_i$ and $\hat{\beta}_i$ $(i = X, D, Y)$ solve the following optimization

$$(\hat{\alpha}_X, \hat{\beta}_X) = argmin_{\alpha,\beta} \sum_{i=1}^{n} (X_{1i} - \alpha - \sum_{0\le|u|\le 1} \beta^{\intercal}(Z_i - z)^u)^2 K_g(Z_i - z)\,, \tag{11}$$

$$(\hat{\alpha}_Y, \hat{\beta}_Y) = argmin_{\alpha,\beta} \sum_{i=1}^{n} (Y_i - \alpha - \sum_{0\le|u|\le p} \beta^{\intercal}(X_i - x, \hat{V}_i - v)^u)^2 K_h((X_i - x, \hat{V}_i - v))\,, \tag{12}$$

$$(\hat{\alpha}_D, \hat{\beta}_D) = argmin_{\alpha,\beta} \sum_{i=1}^{n} (D_i - \alpha - \sum_{0\le|u|\le p} \beta^{\intercal}(X_i - x, \hat{V}_i - v)^u)^2 K_h((X_i - x, \hat{V}_i - v))\,. \tag{13}$$

respectively, where $\sum_{0\le|u|\le p}$ denotes the summation over all nonnegative integer vector $u = (u_1, \cdots, u_{d_u})$ with $|u| = \sum_{l=1}^{d_u} u_l$. Also, for any vector $w = (w_1, \cdots, w_{d_u})$, $w^u = (w_1^{u_1}, \cdots, w_{d_u}^{u_{d_u}})$, $K_g(w) = g^{-d_u}K(w/g)$, $K_h(w) = h^{-d_u}K(w/h)$. Here, g and h are bandwidths in first and second stage estimation.

By applying (8) and direct calculation, we expand $\hat{\gamma} - \gamma$ as below,

$$\frac{\left\{\hat{\mathbb{E}}\left[\frac{\widehat{\partial^2\mathbb{E}}(Y|X,\hat{V})}{\partial X_1 \partial V}\right] - \mathbb{E}\left[\frac{\partial^2\mathbb{E}(Y|X,V)}{\partial X_1 \partial V}\right]\right\}\mathbb{E}\left[\frac{\partial^2\mathbb{E}(D|X,V)}{\partial X_1 \partial V}\right] - \left\{\hat{\mathbb{E}}\left[\frac{\widehat{\partial^2\mathbb{E}}(D|X,\hat{V})}{\partial X_1 \partial V}\right] - \mathbb{E}\left[\frac{\partial^2\mathbb{E}(D|X,V)}{\partial X_1 \partial \hat{V}}\right]\right\}\mathbb{E}\left[\frac{\partial^2\mathbb{E}(Y|X,V)}{\partial X_1 \partial V}\right]}{\hat{\mathbb{E}}\left[\frac{\widehat{\partial^2\mathbb{E}}(D|X,\hat{V})}{\partial X_1 \partial V}\right]\mathbb{E}\left[\frac{\partial^2\mathbb{E}(D|X,V)}{\partial X_1 \partial V}\right]}\,. \tag{14}$$

To show that $\hat{\gamma}$ converges to γ at the parametric rate, the major step is to demonstrate that

$$\sqrt{n}\left\{\hat{\mathbb{E}}\left[\frac{\widehat{\partial^2\mathbb{E}}(Y|X,\hat{V})}{\partial X_1 \partial V}\right] - \mathbb{E}\left[\frac{\partial^2\mathbb{E}(Y|X,V)}{\partial X_1 \partial V}\right]\right\} = O_P(1)$$

[1] $\mathbf{N}$ is given by (27) in the Appendix.

and also $\sqrt{n}\left\{\hat{\mathbb{E}}\left[\frac{\widehat{\partial^2\mathbb{E}}(D|X,\hat{V})}{\partial X_1 \partial V}\right] - \mathbb{E}\left[\frac{\partial^2\mathbb{E}(D|X,V)}{\partial X_1 \partial V}\right]\right\} = O_P(1)$. As it's easy to see that the two identities can be proved by the same method, to save space, we only state the result for the first one here in Proposition 1. The result for the other is presented in the Appendix as Corollary A.1.

To apply empirical process methods, we need to restrict the complexity of the smooth class that the cross-partial derivative of the expectation belongs to. We use the smooth class defined in van der Vaart and Wellner (1996).

Definition $C_M^\alpha(\mathcal{S})$**:** For any vector $s = (s_1, \cdots, s_{d_s})$ in a compact set $\mathcal{S} \in \mathcal{R}^{d_s}$ and any integer vector $q = (q_1, \cdots, q_d)$, let D^q denote the differential operator $D^q = \frac{\partial^{|q|}}{\partial}$, where $|q| = \sum_{l=1}^d q_l$. Let $\underline{\alpha}$ be the greatest integer smaller than α. Define

$$\|g\|_\alpha = \max_{|q|\leq\underline{\alpha}} \sup |D^q g(s)| + \max_{|q|\leq\underline{\alpha}} \sup_{s\neq s'} |D^q g(s) - D^q g(s')|/\|s - s'\|^{\alpha-\underline{\alpha}},$$

where the supremum is taken over the interior of $\mathcal{S}$. Then $C_M^\alpha(\mathcal{S})$ is the set of all continuous functions $g : \mathcal{S} \mapsto \mathcal{R}$ with $\|g\|_\alpha \leq M$.

Before we state our results, we impose the following assumptions.

Assumption 6 *For bandwidth g and h, as $n \to \infty$, we have (i)$\frac{\log n}{\sqrt{n}g^{d_z}} \to 0$ and $\sqrt{n}g^4 \to 0$; (ii) $nh^{2p} \to 0$ and $\frac{nh^{d+5}}{\log n} \to \infty$. (iii) $ng^{d_z}h^{d+7} \to \infty$.*

Assumption 7 *$\frac{\partial^2\mathbb{E}(Y|X,V)}{\partial X_1 \partial V}, \frac{\partial^2\mathbb{E}(D|X,V)}{\partial X_1 \partial V} \in C_M^\alpha(\mathcal{X} \times \mathcal{V})$, where $\underline{\alpha} = \max\{2, (d+1)/2\}$ for odd d and $\underline{\alpha} = \max\{2, d/2\}$ for even d.*

Assumption 8 *$\mathbb{E}[\exp(l|\xi|)|X, V] \leq C$ almost surely for a constant $C > 0$ and $l > 0$ small enough.*

Assumption 6 is the conditions on bandwidths h and g. The second part of this assumption is comparable to the corresponding assumption in Li et al. (2003). The first and third part are due to reasons similar to the ones in Corollary 6 in Mammen et al. (2012a). Assumption 6(i) is about the first stage local linear estimation, $\sqrt{n}g^4 \to 0$ part can be relaxed if a higher order kernel is used. Assumption 7 and Assumption 8 are also used by Mammen et al. (2012a) to apply empirical process theory. Assumption 7 is for the stochastic equicontinuity argument in empirical process theory. It implies $\frac{\partial^2\mathbb{E}(Y|X,V)}{\partial X_1 \partial V}$ and $\frac{\partial^2\mathbb{E}(D|X,V)}{\partial X_1 \partial V}$ belong to a Donsker class by Example 19.9 in van der Vaart (2000). The smooth class is not limited to the one we choose, and this assumption can be replaced by imposing condition on the bracket entropy of the class directly. Assumption 6 and 7 together will ensure the Accuracy and Complexity

Assumptions in Mammen et al. (2012a) are satisfied, which enables some of their results are applicable to our model.

Let $m(x,v)=\mathbb{E}(Y|X=x,V=v)$. For any matrix A, $[A]_{i,j}$ represents the (i,j) entry of matrix A. We have the following result with the generated regressor $\hat{V}$,

Proposition 1: Under Assumptions 1-8, we have

$$\sqrt{n}\left\{\hat{\mathbb{E}}\left[\frac{\widehat{\partial^2\mathbb{E}}(Y|X,\hat{V})}{\partial X_1\partial V}\right]-\mathbb{E}\left[\frac{\partial^2\mathbb{E}(Y|X,V)}{\partial X_1\partial V}\right]-h^{p-1}\text{Bias}_Y\right\}\xrightarrow{d}N(0,\Lambda)\,,$$

where

$$\begin{aligned}\Lambda=&\,4\mathbb{E}\{\sigma_\xi^2[R(X,V)]_{t,1}^2\}+\mathbb{E}\left\{V^2\mathbb{E}_{X|Z}\left[\frac{\partial^3\mathbb{E}(Y|X,V)}{\partial X_1\partial V^2}\right]^2\right\}+\mathbb{E}\left[\frac{\partial^2\mathbb{E}(Y|X,V)}{\partial X_1\partial V}\right]^2\\&-2\mathbb{E}\left\{V\mathbb{E}_{X|Z}\left[\frac{\partial^3\mathbb{E}(Y|X,V)}{\partial X_1\partial V^2}\right]\frac{\partial^2\mathbb{E}(Y|X,V)}{\partial X_1\partial V}\right\},\\\text{Bias}_Y=&\,2e_t^\intercal M^{-1}B\mathbb{E}[m^{(p+1)}(X,V)]\,,\end{aligned}$$

M, B, $R(x,v)$ and $m_{p+1}(x,v)$ are defined by (28), (29), (31) and the line above (31) respectively in the Appendix.

As we can see, this result is about the estimation of the average of second order derivatives. In Li et al. (2003), they show that the estimator of the average of first order derivatives converges at parametric rate. In Proposition 1, we obtain the $\sqrt{n}$-consistent result for the estimator of the average of second order derivatives as a by-product, which further extends the exiting literature.

Now, we introduce our main result that $\hat{\gamma}$ converges at the parametric rate,

Theorem 1: Under Assumptions 1-8, we have

$$\sqrt{n}(\hat{\gamma}-\gamma)\xrightarrow{d}N\left(0,\frac{1}{\mathbb{E}\left[\frac{\partial^2\mathbb{E}(D|X,V)}{\partial X_1\partial V}\right]^2}4\sigma_\varsigma^2\mathbb{E}\{[R(X,V)]_{t,1}^2\}\right)\,,\tag{15}$$

where $t=2d+3$, σ_ς^2 is the variance of $\varsigma=\xi-\gamma u$ and $[R(x,v)]_{t,1}$ is the $(t,1)$ entity in the matrix $R(x,v)$ defined by equation (31) in the Appendix.

The result indicates that our estimator no longer has biased terms from the local polynomial regression. This desirable property comes from the special ratio form of the estimator we proposed and also the equality (8), as the bias terms from the estimation of the denominator and numerator are exactly

cancelled. The variance depends on both disturbances in the regression of Y and D. From the result, we see that the larger γ, the smaller the variance, which is consistent with the intuition. Also, $R(X,V)$ depends on the joint density of (X,V) and its second derivatives.

The estimation procedure and asymptotic results for case (i) is much simpler. For case (i), D is the only endogenous variable and we consider V in the model an observed variable. Then, the model becomes $Y = \lambda(X)+\gamma D+\rho(V)+\varsigma$. To estimate γ, we no longer need the first step in the aforementioned procedure and the estimator in the second step is obtained by replacing $\hat{V}$ by V in (10). Therefore, we have

$$\hat{\gamma} = \hat{\mathbb{E}}\left[\frac{\widehat{\partial^2\mathbb{E}}(Y|X,V)}{\partial X_1 \partial V}\right] \Big/ \hat{\mathbb{E}}\left[\frac{\widehat{\partial^2\mathbb{E}}(D|X,V)}{\partial X_1 \partial V}\right].$$

Applying the result from the proof Proposition 1 and also using similar method of Theorem 1, we obtain the following Corollary:

Corollary 1: Under Assumptions 1-8, we have the asymptotic distribution of $\hat{\gamma}$ is given by (15).

In Theorem 1, we find that the generated regressor has no influence on the estimation asymptotically as its effects in $\frac{\widehat{\partial^2\mathbb{E}}(Y|X,\hat{V})}{\partial X_1 \partial V}$ and $\frac{\widehat{\partial^2\mathbb{E}}(D|X,\hat{V})}{\partial X_1 \partial V}$ are exactly cancelled by equality (8). Therefore, the results are identical with or without the generated regressor for both cases.

By (9), $\lambda(\cdot)$ and $\rho(\cdot)$ can be estimated by regressing $Y - \hat{\gamma}D$ on $X, \hat{V}$ as in Mammen et al. (2012a). As $\hat{\gamma}$ converges at parametric rate, the result in Mammen et al. (2012a) holds. Therefore, we can apply Mammen et al. (2012a) directly.

4 Extensions of the Model:

4.1 Linear Controls:

We expect that in most empirical applications of our methods, the variables of interest will be X_1 and D (and perhaps V), and Z_1 will represent covariates or controls that the researchers will use to argue the validity of the instrument Z_2 more credibly. In a typical realistic empirical application we expect that the number of controls, i.e. the dimension of Z_1 to be large. In such cases, estimating $\frac{\partial^2\mathbb{E}(Y|X_1,Z_1,V)}{\partial X_1 \partial V}$ and $\frac{\partial^2\mathbb{E}(D|X_1,Z_1,V)}{\partial X_1 \partial V}$ nonparametrically will be challenging due to data constraints. To make our methods more readily accessible for applied researchers, in this section we discuss how the model can be estimated if

the controls are added linearly, that is if $\lambda(X_1, Z_1) = g(X_1) + Z_1^\intercal\theta$. Then our model becomes,

$$Y = g(X_1) + D\gamma + Z_1^\intercal\theta + \varepsilon$$
$$X_1 = \pi(Z) + V\,.$$

In this case, under Assumptions 1 and 4, we can identify γ exactly the same as before. Once γ is identified, we can define

$$\tilde{Y} = Y - D\gamma.$$

Then θ is identified using Robinson (1988), since

$$\tilde{Y} - \mathbb{E}(\tilde{Y}|X_1, V) = [Z_1 - \mathbb{E}(Z_1|X_1, V)]^\intercal\theta + e_1\,.$$

This method still requires nonparametric estimation of $\frac{\partial^2\mathbb{E}(\tilde{Y}|X_1,V,Z_1)}{\partial X_1\partial V}$ and $\frac{\partial^2\mathbb{E}(D|X_1,V,Z_1)}{\partial X_1\partial V}$, which could be challenging if dimension of Z_1 is large. Alternatively, under the assumption that $\mathbb{E}(\varepsilon|X_1, V) = \mathbb{E}(\varepsilon|V)$, which is implied by $\mathbb{E}(\varepsilon|X_1, V, Z_1) = \mathbb{E}(\varepsilon|V)$, we get

$$\frac{\partial^2\mathbb{E}(Y|X_1,V)}{\partial X_1\partial V} = \frac{\partial^2\mathbb{E}(D|X_1,V)}{\partial X_1\partial V}\gamma + \frac{\partial^2\mathbb{E}(Z_1|X_1,V)}{\partial X_1\partial V}^\intercal\theta.$$

Then if $\left(\frac{\partial^2\mathbb{E}(D|X_1,V)}{\partial X_1\partial V}, \frac{\partial^2\mathbb{E}(Z_1|X_1,V)}{\partial X_1\partial V}^\intercal\right)$ are linearly independent, $(\gamma, \theta^\intercal)^\intercal$ is identified. In that case we can estimate $(\gamma, \theta^\intercal)^\intercal$ by regressing the estimated values of $\frac{\partial^2\mathbb{E}(Y|X_1,V)}{\partial X_1\partial V}$ on estimated values of $\frac{\partial^2\mathbb{E}(D|X_1,V)}{\partial X_1\partial V}$ and estimated values of $\frac{\partial^2\mathbb{E}(Z_1|X_1,V)}{\partial X_1\partial V}$.

4.2 Nonlinear Functions of D:

It is easy to see that polynomials of D and Z_1 can be incorporated into the model in a straightforward way. Here to keep notation simple, we assume the dimensions of X_1, D, Z_1 are all 1, and we write X to denote X_1. Consider

$$Y = g(X) + \sum_{j=0}^{K} D^j Z_1^{K-j}\gamma_j + \varepsilon.$$

As long as $\mathbb{E}(\varepsilon|X,V) = \mathbb{E}(\varepsilon|V)$, and $\mathbb{E}(S^\intercal S)$ is full rank, where

$$S = \left[\frac{\partial^2 \mathbb{E}(Z_1^K|X,V)}{\partial X \partial V}, \frac{\partial^2 \mathbb{E}(Z_1^{K-1}D|X,V)}{\partial X \partial V}, \cdots, \frac{\partial^2 \mathbb{E}(D^K|X,V)}{\partial X \partial V}\right],$$

we can identify γ, g and ρ, where $\gamma = (\gamma_0, \cdots, \gamma_K)^\intercal$. Similarly, the model

$$Y = g(X) + \delta(D, Z_1; \theta_0) + \varepsilon,$$

where the function δ is known up to a finite dimensional parameter vector θ can be estimated by minimizing

$$\left[\frac{\partial^2 \mathbb{E}(Y|X,V)}{\partial X \partial V} - \frac{\partial^2 \mathbb{E}(\delta(D, Z_1; \theta)|X,V)}{\partial X \partial V}\right]^2$$

with respect to θ.

Finally, it might be possible to extend our methods to the nonparametric $\delta(D, Z_1)$ case. Investigation of this question is left for future research.

4.3 Interactions between X and D:

In this section we consider how interactions between X and D could be incorporated into the basic model. For ease of notation, we suppress Z_1, and also assume that the dimensions of $X = X_1$ and D are both 1. So we consider the model

$$Y = g(X) + \sum_{j=1}^{K} D^j X^{K-j} \gamma_j + \varepsilon.$$

As long as $\mathbb{E}(\varepsilon|X,V) = \mathbb{E}(\varepsilon|V)$, and $\mathbb{E}(SS^\intercal)$ is full rank, where $S$is the row vector whose j^{th} component equals

$$m_{XV}^D(X,V)X^{K-j} + m_V^D(X,V)(K-j)X^{K-j-1},$$

with $m^D(X,V)$, m_V^D, and m_{XV}^D denoting $E(D|X,V)$, its partial derivative with respect to D and its cross partial derivative with respect to X and V, respectively, we can identify γ, g and ρ, where $\gamma = (\gamma_1, \cdots, \gamma_K)^\intercal$.

This extension suggests that it might be possible to extend our method to the analysis of

$$Y = \theta(X, D) + \varepsilon$$

as well. We leave answering this question to future research as well.

5 Finite Sample Performance

To evaluate finite sample performance of our estimator we did a Monte Carlo study. In this section we report the results of this study. We consider two basic specifications. The first specification satisfies Assumption 1, and the second one does not. Moreover, in the second specification we investigate the performance of our estimator for mild and severe violations of Assumption 1. The specifications we consider only differ in the form of the outcome equation. The outcome equations we consider are all of the form

$$Y \quad = \quad 0.5X + D + \kappa D\varsigma + \varsigma \,. \tag{16}$$

In DGP1, $\kappa = 0$, so D and ς do not interact, even though they are correlated. In DGP2 and DGP3 $\kappa = 0.01$ and $\kappa = 1$, respectively. For all cases, D, X, Z and V are as follows:

$$\begin{aligned} D &= 1.5XV + \varsigma_1 + \varsigma_2 + u \,, \\ \varsigma &= 2\varsigma_1 + 2V^2 \,, \\ X &= 0.3 - 1.5Z - 0.5Z^2 + V \,. \end{aligned}$$

where V, u are independent $N(0,1)$ random variables, and Z serves as an IV for X. Moreover, $\varsigma_1, \varsigma_2, \varsigma_3$ and Z, are all drawn from $N(0, 0.5)$. They are independent from each other and from both u and V. In order to compare our estimator with both the conventional OLS and IV regression, construct $Z_D = \varsigma_2 + \varsigma_3$ to be the additional instrumental variable for D required in the IV regression. To implement our method, we use product kernel and choose normal density function as our kernel function, i.e., $K(u) = \frac{1}{\sqrt{2\pi}} \exp^{-\frac{u^2}{2}}$. In our simulation, we use Silverman's rule of thumb for the first stage bandwidth, $g = 1.06\sigma(Z)N^{-1/5}$ and choose $h = 2$ for both X and V dimensions in the second stage bandwidth.[2] The simulations reported here consider sample sizes $N = 250, 500, 1000$, and are repeated 500 times. Summary statistics reported

[2] We also tried $h = 1.5$ and $h = 2.5$ and the results were similar.

include the sample mean (MEAN), standard deviation (SD), and root-mean-square error (RMSE), as well as the median (MED) and median absolute error (MAE). Tables 1, 2 and 3 summarize our simulation results for $\kappa = 0, 0.01, 1$, respectively. Moreover, in each case, γ is estimated four different ways: using OLS, our estimator (HKY), using the efficient IV estimation which uses the constant, Z and Z_D as intruments (IV-spec-I), and 2SLS using Z as an instrument for X and treating D as exogenous (IV-spec-II).

When $\kappa = 0$, both HKY estimator and IV-spec-I estimator are consistent, but OLS and IV-spec-II estimators are not. We see the bias in OLS and IV-spec-II estimators in Table 1. It is notable that our estimator outperforms the IV estimator with two valid IV's. When $\kappa = 0.01$ and when $\kappa = 1$, none of the estimators are consistent. When $\kappa = 0.01$ both IV-spec-I and HKY estimators seem to perform reasonably well. When $\kappa = 1$ all of the estimators seem do poorly. While the IV-spec-I estimator has the lowest bias among the estimators we considered, the HKY estimator seems to have lower RMSE.

Table 1: Finite Sample Performance of OLS, IV and HKY with $\gamma = 1$ when Assumption 1 holds

Estimators	MEAN	SD	RMSE	MED	MAE
		$N = 250$			
OLS	1.752	0.079	0.079	1.750	0.063
IV spec-I	0.754	1.308	1.307	0.983	0.564
IV spec-II	1.753	0.079	0.078	1.751	0.063
HKY	0.949	0.221	0.221	0.968	0.174
		$N = 500$			
OLS	1.749	0.054	0.054	1.748	0.042
IV spec-I	0.921	0.489	0.489	1.011	0.276
IV spec-II	1.750	0.054	0.054	1.749	0.042
HKY	0.970	0.136	0.136	0.973	0.108
		$N = 1000$			
OLS	1.752	0.040	0.040	1.753	0.033
IV spec-I	0.956	0.254	0.254	0.988	0.173
IV spec-II	1.752	0.040	0.040	1.753	0.032
HKY	0.978	0.102	0.102	0.978	0.082

Table 2: Finite Sample Performance of OLS, IV and HKY with $\gamma = 1$ when $\kappa = 0.01$

Estimators	MEAN	SD	RMSE	MED	MAE
		$N = 250$			
OLS	1.829	0.090	0.090	1.827	0.072
IV spec-I	0.756	1.430	1.429	1.009	0.608
IV spec-II	1.831	0.090	0.089	1.828	0.072
HKY	0.996	0.224	0.224	1.016	0.176
		$N = 500$			
OLS	1.829	0.062	0.062	1.828	0.049
IV spec-I	0.934	0.527	0.526	1.031	0.293
IV spec-II	1.830	0.062	0.062	1.829	0.049
HKY	1.017	0.138	0.138	1.019	0.110
		$N = 1000$			
OLS	1.834	0.048	0.048	1.835	0.038
IV spec-I	0.973	0.272	0.272	1.009	0.185
IV spec-II	1.834	0.047	0.047	1.834	0.038
HKY	1.026	0.104	0.104	1.024	0.083

Table 3: Finite Sample Performance of OLS, IV and HKY with $\gamma = 1$ when $\kappa = 1$

Estimators	MEAN	SD	RMSE	MED	MAE
		$N = 250$			
OLS	9.512	1.958	1.956	9.264	1.469
IV spec-I	0.9560	14.525	14.511	3.176	5.445
IV spec-II	9.541	1.976	1.974	9.295	1.478
HKY	5.633	1.227	1.226	5.605	0.974
		$N = 500$			
OLS	9.749	1.588	1.587	9.506	1.226
IV spec-I	2.248	4.861	4.856	2.939	2.563
IV spec-II	9.771	1.598	1.596	9.513	1.231
HKY	5.660	0.869	0.868	5.631	0.683
		$N = 1000$			
OLS	9.891	1.193	1.192	9.789	0.918
IV spec-I	2.559	2.401	2.399	2.950	1.707
IV spec-II	9.901	1.197	1.196	9.793	0.920
HKY	5.741	0.603	0.602	5.742	0.471

6 Empirical Example

6.1 Background

Smoking during pregnancy and its causal effect on birth outcomes has been a long standing issue in both medical and policy circles. The most comprehensive treatment of the question is provided by Almond et al. (2005), under a selection-on-observables framework, and estimates a significant difference in birth weights of around 200 grams between smokers and non-smokers. However, as mentioned earlier, the primary concern in the causal estimation of the effect of smoking on birth weight remains due to the lack of an exogenous source of variation. As Caetano (2015) argues even after using the most comprehensive set of controls, as in Almond et al. (2005), there is significant left over selection that hinders meaningful causal inference. The framework developed in this paper can thus, potentially help contribute to this persistent problem in the literature.

To formulate the empirical setup more precisely, consider the question explored by Currie and Moretti (2003) (henceforth, CM): the effect of maternal education on birth outcomes. Their setup can be expressed in terms of the standard case of one endogenous variable, mother's education attainment, and one instrumental variable, the number of colleges in the county where the mother was resident at age 17.[3] Applying this setup to our question we can have two potentially endogenous variables of interest in our structural equation: maternal education and maternal smoking, and an IV for the former. Thus using the framework developed so far in this paper, we can use the artificially generated exogenous source of variation to estimate the causal effect of smoking on birth weight.

6.2 Data

We use Vital Statistics data from 1993-2002 and closely follow the sample restrictions used by CM. In particular, we restrict our sample to all singleton births to White mothers aged between 24 and 45 years. We also drop mothers that reside in counties with less than 100,000 population. This leaves us with a sample size of close to 3.5 million births. The key variables of interest for us are the two endogenous variables in our framework, the number of years of schooling of the mother, X, and the reported number of cigarettes smoked by the mother during pregnancy, D. Following the previous literature, our key birth outcome of interest is the birth weight of an infant in grams, Y. Our instrument, Z, is the total number

[3] In their paper, CM actually include separate IVs for the number of four-year colleges and the number of two-year colleges in the county of the mother at age 17.

of four-year and two-year colleges in the county where the mother resides at age 17.[4]

6.3 Estimation

The first step of the estimation regresses maternal education, X, on our IV while controlling for maternal age dummies, 10-year birth of cohort dummies and mother's county-at-age-17*year of birth fixed effects. The residual from this regression, $\hat{V}$, gives us the variable that is analogous to a control function in our framework. In the next step, we estimate conditional expectations of both Y and D given X and $\hat{V}$ using a third order polynomial.[5] Using the coefficients on the interaction terms from the above two series regressions we can then construct, $\frac{\partial^2 \mathbb{E}(Y|X,\hat{V})}{\partial X \partial V}$ and $\frac{\partial^2 \mathbb{E}(D|X,\hat{V})}{\partial X \partial V}$, respectively. The final step then regresses $\frac{\partial^2 \mathbb{E}(Y|X,\hat{V})}{\partial X \partial V}$ on $\frac{\partial^2 \mathbb{E}(D|X,\hat{V})}{\partial X \partial V}$ to estimate $\hat{\gamma}$, the effect of maternal smoking on birth weight.

6.4 Results

This subsection presents the results from the above outlined procedure to estimate the causal effect of smoking during pregnancy on birth weight. The first column of Table 4 reports OLS results using the basic set of controls described above and estimates an effect of -17.46 grams per cigarette smoked during pregnancy.[6] Specification - II then uses a much more detailed set of covariates including demographics of the parents, pregnancy characteristics, pregnancy history, and various interactions between them. This specification closely follows the most comprehensive one used in the literature so far by Almond et al. (2005). The results, however, are fairly consistent across the first two columns.

The last two columns finally present results from the methodology developed in this paper, implementing the estimation details given in the above section. Using a 3rd order series estimator we find an effect size of -29.70 grams per cigarette smoked, which is significantly larger in magnitude compared to the OLS estimates. The last column presents results from a 4th order series estimator with the effect size being slightly smaller compared to column 3.

One might expect the OLS estimates to actually decrease in magnitude given that mother's who smoke are more likely to be selected negatively on unobservables. However, our estimation methodology is inherently instrument variable based, and hence recovers only a local average treatment effect (LATE). Using terminology from the treatment effects literature, the group of 'compliers' for our IV are mothers

[4]We are extremely grateful to Janet Currie and Enrico Moretti for graciously providing us with their novel dataset on college openings in the US.

[5]This birth weight variable is cleaned off the variation solely due to maternal age dummies, 10-year birth cohort dummies and mother's county-at-age-17*year of birth fixed effects.

[6]We follow CM closely for this specification.

Table 4: Effect of Maternal Smoking on Birth Weight

	OLS		Series Estimator	
	Spec - I	Spec - II	Order - 3	Order - 4
Cigarettes Smoked	−17.46**	−16.14**	−29.70**	−27.40**
	(0.202)	(0.191)	(1.493)	(0.007)
Years of Schooling	14.95**	9.020**	–	–
	(0.499)	(0.576)		
Number of Observations	3,443,755	3,443,755	2,824,880	2,824,880

**,* Indicates significance at 1, and 5 percent, respectively. OLS specification - I includes non-parametric controls for mother's age, dummies for 10-year birth cohort and mother's county-at-age-17*year of birth fixed effect. Specification - II controls for a more elaborate set of controls with various interactions and closely follows the one used by Almond et al. (2005). The last two columns then present estimates using our methodology and employ a series estimator of order 3 and 4, respectively. Standard errors are clustered at the mother's county of residence at age 17 level.

who are more likely to go to college as a result of more college openings in their county of residence. These mothers, in turn, are also more likely to be positively selected on other unobservable dimensions which could be positively correlated with mothering ability. This, therefore, can account for the increase in magnitude in effect sizes that we document using our methodology compared to the OLS estimates.

7 Conclusion

In this paper, we proposed a novel identification method for a partially linear model with multiple endogenous variables of interest in which the number of available excluded instruments is strictly less than the number of endogenous variables. One of the endogenous variables in the model is assumed to be continuous and its structural relation to the outcome of interest could be nonparametric. We assume that there is one excluded instrument that could even be discrete for this endogenous variable. Using this instrument we identify a control function for the continuous endogenous regressor. Inclusion of this control function does not necessarily control for the endogeneity in other regressors. Nevertheless, we can still identify all parameters of interest. In particular, given the partially linear structure of the model, the change in change in the conditional outcome given the continuous endogenous variable and its control function with respect to a change in the continuous endogenous variable first and its control function second can only be the result of a corresponding change in change in the conditional expectation of the other endogenous variables given the continuous endogenous variable and its control function times the coefficient vector on the other endogenous regressors. This way, we can identify this coefficient. After identifying this coefficient, we discuss how other parameters of interest can be identified in this model. We

also provide an easy-to-compute $\sqrt{n}$-consistent estimator for the coefficients on the endogenous variables for which no excluded instruments are available. We show that the proposed estimator is $\sqrt{n}$-normal. A by-product of our method is an $\sqrt{n}$-consistent estimator of the average of second order derivatives, which is also a new result in the literature to our knowledge. The Monte Carlo simulation results demonstrate our estimator has fairly good finite sample performance. In the empirical section, we use the methods proposed in this paper to assess the causal impact of mother's smoking during pregnancy on baby's birth weight. Our results seem to be in line with findings of the previous literature. We also outline how our identification method can be used in extensions of the basic model we consider. We leave detailed analysis of these extensions for future research.

References

[1] Ai, C. and X. Chen (2003). Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions. *Econometrica* **71** (6), 1795-1843.

[2] Almond Douglas, Kenneth Y. Chay and David S. Lee (2005). The Costs of Low Birth Weight. *Quarterly Journal of Economics* **120** (3), 1031-1083.

[3] Blundell, R. W. and J.L. Powell (2003). Endogeneity in Nonparametric and Semiparametric Regression Models. *Advances in Economics and Econometrics: Theory and Applications*, Eighth World Congress, Vol. II. Cambridge: Cambridge University Press.

[4] Blundell, R. and J. Powell (2004). Endogeneity in Semiparametric Binary Response Models. *Review of Economic Studies* **71**, 655-679.

[5] Caetano, Carolina (2015). A Test of Endogeneity without Instrumental Variables in Models with Bunching. *Econometrica*, *forthcoming*.

[6] Chen, X. and D. Pouzo (2009). Efficient Estimation of Semiparametric Conditional Moment Models with Possibly Nonsmooth Residuals. *Journal of Econometrics* **152**, 46-60.

[7] Currie, Janet and Enrico Moretti (2003). Mother's Education and the Intergenerational Transmission of Human Capital: Evidence from College Openings. *Quarterly Journal of Economics* **118** (4), 1495-1532.

[8] Escanciano, J.C., D.T. Jacho-Chávez and A. Lewbel (2012). Identification and Estimation of Semiparametric Two Step Models. *Unpublished manuscript*.

[9] Escanciano, J.C., D.T. Jacho-Chávez, and A. Lewbel (2014). Uniform Convergence of Weighted Sums of Non and Semiparametric Residuals for Estimation and Testing. *Journal of Econometrics* **178** (P3), pages 426-443.

[10] Florens, J.-P. (2003). Inverse Problems and Structural Econometrics: the Example of Instrumental Variables. *Advances in Economics and Econometrics: Theory and Applications*, Eight World Congress, Econometric Society Monograph Series, ESM 36, Volume II, 284-312. Cambridge: Cambridge University Press.

[11] Florens, J.-P., J. Johannes and S. Van Bellegem (2012). Instrumental Regression in Partially Linear Models. *The Econometrics Journal* **15** (2), 304-324.

[12] Hahn, J. and G. Ridder (2013). The Asymptotic Variance of Semi-parametric Estimators with Generated Regressors. *Econometrica* **81** (1), 315-340.

[13] Heckman, J. J., H. Ichimura, and P. Todd (1998, April). Matching as an Econometric Evaluation Estimator. *Review of Economic Studies* **65** (2), 261-294.

[14] Imbens, G. W. and W.K. Newey (2009). Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity. *Econometrica* **77**, 1481-1512.

[15] Jacob, B., L. Lefgren and E. Moretti (2007). The Dynamics of Criminal Behavior: Evidence from Weather Shocks *The Journal of Human Resources* **42** (3), 489-527

[16] Kramer (1987). Determinants of Low Birth Weight: Methodological Assessment and Meta-analysis. *Bull World Health Organ* **65**, 663-737.

[17] Lee (2013). Partial Mean Processes with Generated Regressors: Continuous Treatment Effects and Nonseparable Models. *Working paper*.

[18] Li, Q., Lu and Ullah (2003). Multivariate Local Polynomial Regression for Estimating Average Derivatives. *Journal of Nonparametric Statistics* 15:4-5, 607-624.

[19] Mammen, E., C. Rothe, and M. Schienle (2012a). Nonparametric Regression with Nonparametrically Generated Covariates. *Annals of Statistics* **40**, 1132-1170.

[20] Mammen, E., C. Rothe, and M. Schienle (2012b). Semiparametric Estimation with Generated Covariates, forthcoming in *Econometric Theory*.

[21] Masry, E. (1996a). Multivariate Regression Estimation Local Polynomial Fitting for Time Series. *Stochastic Processes and Their Applications* **65**, 81-101.

[22] Masry, E. (1996b). Multivariate Local Polynomial Regression for Time Series: Uniform Strong Consistency and Rates. *Journal of Time Series Analysis* **17**, 571-599.

[23] Newey, W.K., J.L. Powell and F. Vella (1999). Nonparametric Estimation of Triangular Simultaneous Equations Models. *Econometrica* **67**, 565-603.

[24] Robinson, P.M. (1988). Root-N-consistent Semiparametric Regression. *Econometrica* **56** (4), 931-954.

[25] Speckman, P. (1988). Kernel Smoothing in Partial Linear Models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 413-436.

[26] Tominey, Emma (2007). Maternal Smoking during Pregnancy and Early Child Outcomes. *CEP Discussion Papers*. Centre for Economic Performance, LSE.

[27] van der Vaart, A. (2000). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

[28] van der Vaart, A., and J. Wellner (1996). *Weak Convergence and Empirical Processes: with Applications to Statistics*. Springer Verlag.

A Mathematical Proof

We first introduce the definition of empirical measure that is commonly used in empirical process theory. For any $X_1, \cdots, X_n$ that are i.i.d. random variables with distribution P and for any measurable function f, the empirical measure G_n is defined as

$$f \mapsto G_n f = \sqrt{n}(P_n - P)f = \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f\right]. \tag{17}$$

We first prove Proposition 1, which indicates that $\sqrt{n}\left\{\hat{\mathbb{E}}\left[\frac{\widehat{\partial^2\mathbb{E}}(Y|X,\hat{V})}{\partial X_1 \partial V}\right] - \mathbb{E}\left[\frac{\partial^2\mathbb{E}(Y|X,V)}{\partial X_1 \partial V}\right]\right\} = O_P(1)$. Similarly, we have Corollary A.1 which states that $\sqrt{n}\left\{\hat{\mathbb{E}}\left[\frac{\widehat{\partial^2\mathbb{E}}(D|X,\hat{V})}{\partial X_1 \partial V}\right] - \mathbb{E}\left[\frac{\partial^2\mathbb{E}(D|X,V)}{\partial X_1 \partial V}\right]\right\} = O_P(1)$. Finally, we apply Delta method on (14) to obtain the results of Theorem 1.

Proof of Proposition 1: We use the directional derivative to decompose the influences of the generated regressor. Following the proof of Corollary 6 in Mammen et al. (2012a), let

$$\bar{f}(x,z) = (\bar{f}_1, \bar{f}_2) = \left(\frac{\partial^2\mathbb{E}[Y|X=x, V=\bar{f}_2(x,z)]}{\partial X_1 \partial V}, x_1 - \pi(z)\right),$$
$$\hat{f}(x,z) = (\hat{f}_1, \hat{f}_2) = \left(\frac{\widehat{\partial^2\mathbb{E}}[Y|X=x, V=\hat{f}_2(x,z)]}{\partial X_1 \partial V}, x_1 - \hat{\pi}(z)\right),$$

where $\bar{f}$ represents the true regression function and $\hat{f}$ is the estimator.

For any $f = (f_1, f_2)$, define

$$S_n(f) := \frac{1}{n}\sum_{i=1}^{n} f_1(X_i, f_2(X_i, Z_i)) - \mathbb{E}\left[\frac{\partial^2\mathbb{E}(Y|X,V)}{\partial X_1 \partial V}\right].$$

Then the directional derivative can be written as

$$\begin{aligned}
\dot{S}_n(\bar{f})[f - \bar{f}] &= \lim_{s\to 0}\frac{1}{s}[S_n(\bar{f} + s(f - \bar{f})) - S_n(\bar{f})] \\
&= \lim_{s\to 0}\frac{1}{s}\frac{1}{n}\sum_{i=1}^{n}\{[\bar{f}_1 + s(f_1 - \bar{f}_1)](X_i, \bar{f}_2 + s(f_2 - \bar{f}_2)) - \bar{f}_1(X_i, \bar{f}_2 + s(f_2 - \bar{f}_2)) \\
&\qquad + \bar{f}_1(X_i, \bar{f}_2 + s(f_2 - \bar{f}_2)) - \bar{f}_1(X_i, \bar{f}_2)\} \\
&= \frac{1}{n}\sum_{i=1}^{n}[f_1 - \bar{f}_1](X_i, \bar{f}_2) + \frac{1}{n}\sum_{i=1}^{n}\bar{f}_1^{(v)}(X_i, \bar{f}_2)\cdot[f_2 - \bar{f}_2](X_i, Z_i) \\
&=: T_{1,n}(f) + T_{2,n}(f),
\end{aligned} \tag{18}$$

where $f_1^{(v)}(x,v) = \partial_v f_1(x,v)$ is the partial derivative of f_1 with respect to v.

For any $f = (f_1, f_2)$ with bounded second derivatives, we have that

$$\begin{aligned}
&\|S_n(f) - S_n(\bar{f}) - \dot{S}_n(\bar{f})[f - \bar{f}]\|_\infty \\
&\qquad = O(\|f_2 - \bar{f}_2\|_\infty^2) + O(\|f_2 - \bar{f}_2\|_\infty \|f_1^{(v)} - \bar{f}_1^{(v)}\|_\infty) =: \text{s.o.}(f)\,. \qquad (19)
\end{aligned}$$

where s.o.(f) defined above is later shown to be of small order. Replacing f in (19) by the estimator $\hat{f}$, we have

$$\begin{aligned}
S_n(\hat{f}) &= \hat{\mathbb{E}}\left[\widehat{\frac{\partial^2 \mathbb{E}}{\partial X_1 \partial V}}(Y|X,\hat{V})\right] - \mathbb{E}\left[\frac{\partial^2 \mathbb{E}(Y|X,V)}{\partial X_1 \partial V}\right] \qquad (20)\\
&= S_n(\bar{f}) + T_{1,n}(\hat{f}) + T_{2,n}(\hat{f}) + \text{s.o.}(\hat{f}) \\
&= \frac{1}{n}\sum_{i=1}^n \bar{f}_1(X_i, \bar{f}_2) - \mathbb{E}\left[\frac{\partial^2 \mathbb{E}(Y|X,V)}{\partial X_1 \partial V}\right] + \frac{1}{n}\sum_{i=1}^n [\hat{f}_1 - \bar{f}_1](X_i, \bar{f}_2) + T_{2,n}(\hat{f}) + \text{s.o.}(\hat{f}) \\
&= \frac{1}{n}\sum_{i=1}^n \hat{f}_1(X_i, \bar{f}_2) - \mathbb{E}\left[\frac{\partial^2 \mathbb{E}(Y|X,V)}{\partial X_1 \partial V}\right] + T_{2,n}(\hat{f}) + \text{s.o.}(\hat{f})\,, \qquad (21)
\end{aligned}$$

where the first term is the averaged estimator for the cross-partial derivative of the conditional expectation with true regressor instead of generated regressor. The rest of proof consists of three parts. Firstly, we use Lemma A.1-A.3 to prove asymptotic normality for the averaged estimator of the cross-partial derivative of the conditional expectation with true regressor V, and we obtain that the estimator for the average second order cross-partial derivatives of the conditional expectation converges at parametric rate as by-product. Secondly, we show that s.o.$(\hat{f}) = o_P(n^{-1/2})$. Thirdly, we apply Lemma A.4 to show $\sqrt{n}T_{2,n}(\hat{f}) = O_p(1)$. Lastly, we combine the results to finish the proof.

Let's start from the first part. With true regressor V, we have

$$\begin{aligned}
\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \widehat{\frac{\partial^2 \mathbb{E}}{\partial X_1 \partial V}}(Y_i|X_i,V_i) - \mathbb{E}\left[\frac{\partial^2 \mathbb{E}(Y|X,V)}{\partial X_1 \partial V}\right]\right) &= G_n\left[\widehat{\frac{\partial^2 \mathbb{E}}{\partial X_1 \partial V}}(Y_i|X_i,V_i) - \frac{\partial^2 \mathbb{E}(Y_i|X_i,V_i)}{\partial X_1 \partial V}\right] \qquad (22)\\
&\quad + G_n\left[\frac{\partial^2 \mathbb{E}(Y_i|X_i,V_i)}{\partial X_1 \partial V}\right] \qquad (23)\\
&\quad + \sqrt{n}\mathbb{E}\left[\widehat{\frac{\partial^2 \mathbb{E}}{\partial X_1 \partial V}}(Y|X,V) - \frac{\partial^2 \mathbb{E}(Y|X,V)}{\partial X_1 \partial V}\right], \qquad (24)
\end{aligned}$$

where G_n is the empirical measure defined in (17). The first term (22) is $o_P(1)$ by the stochastic equicontinuity result in Lemma A.1. The second term (23) is $O_P(1)$ by the Donsker property of $C_M^\alpha(\mathcal{X} \times \mathcal{V})$. The third term (24) contributes to the influence from estimating the cross partial derivatives of the

conditional expectation which we shall see later.

Lemma A.1 (Stochastic Equicontinuity): $G_n\left[\frac{\widehat{\partial^2\mathbb{E}}(Y_i|X_i,V_i)}{\partial X_1\partial V}-\frac{\partial^2\mathbb{E}(Y_i|X_i,V_i)}{\partial X_1\partial V}\right]=o_P(1)$.

Proof of Lemma A.1: Define $Z_{ni}(f)=\frac{1}{\sqrt{n}}f(X_i,V_i)$ indexed by $f\in\mathcal{F}=C_M^\alpha(\mathcal{X}\times\mathcal{V})$. By Assumption 7 and Example 19.9 in van der Vaart (2000), we know that $C_M^\alpha(\mathcal{X}\times\mathcal{V})$ is P-Donsker, and also totally bounded in $L_2(P)$ since for any $\varepsilon>0$, the entropy $\log N_{[\,]}(\varepsilon,C_M^\alpha,L_2(P))<\infty$. The bracketing CLT (van der Vaart and Wellner, 1996) implies that $\sum_{i=1}^n[Z_{ni}(f)-Z_{ni}(\tilde{f})]$ is asymptotic equicontinuous in f with respect to the semimetric $\|f-\tilde{f}\|_2$, which further implies (22) is of small order. We need to verify the conditions of Theorem 2.11.9 in van der Vaart and Wellner (1996).

(i) We know that $|f(x,v)|$ is uniformly bounded as $f\in C_M^\alpha(\mathcal{X}\times\mathcal{V})$. Therefore, for any $\eta>0$, $1_{\|Z_{ni}\|_\mathcal{F}>\eta}=0$ when n is large enough. Obviously, we have $\sum_{i=1}^n\mathbb{E}^*\|Z_{ni}\|_\mathcal{F}1_{\|Z_{ni}\|_\mathcal{F}>\eta}\to 0$ for every $\eta>0$.

(ii) We have $\sum_{i=1}^n\mathbb{E}[Z_{ni}(f)-Z_{ni}(g)]^2=\mathbb{E}[f(X,V)-g(X,V)]^2=o(1)$ for any $\|f-g\|_2=o(1)$.

(iii) $\int_0^{\delta_n}\sqrt{\log N_{[]}(\varepsilon,\mathcal{F},L_2(P))}\,d\varepsilon\to 0$ holds for every $\delta_n\to 0$ also by Assumption 7 and Example 19.9 in van der Vaart (2000) (or Corollary 2.7.2 in van der Vaart and Wellner, 1996). ■

For any integer vector $j=(j_1,\ldots,j_{d+1})$ and random vector $w=(w_1,\cdots,w_{d+1})$, we have already defined that $|j|=j_1+\cdots+j_{d+1}$ and $w^j=(w_1^{j_1},\cdots,w_{d+1}^{j_{d+1}})$. Let $j!=j_1!\times\cdots\times j_{d+1}!$.

For ease of notation, let $W=(X,V)$ and define $m(w)=\mathbb{E}(Y|W=w)$. As in Masry (1996a,b), $k!\hat{b}_k(w)$ estimates the $|k|$th order partial derivative $D^km(w)$ (or $\frac{\partial^{|k|}m(w)}{\partial w_1^{k_1}\cdots\partial w_{d+1}^{k_{d+1}}}$), where $\hat{b}_k(w)$ minimizes the weighted least squares

$$\sum_{i=1}^n\left[Y_i-\sum_{0\le|k|\le p}b_k(w)(W_i-w)^k\right]^2K\left(\frac{W_i-w}{h}\right). \tag{25}$$

Minimizing (25), the F.O.C can be formulated as

$$t_{n,j}=\sum_{0\le|k|\le p}h^{|k|}\hat{b}_k(w)s_{n,j+k}(w),\quad 0\le|j|\le p\,, \tag{26}$$

where

$$t_{n,j} = \frac{1}{n}\sum_{i=1}^{n} Y_i \left(\frac{W_i - w}{h}\right)^j K_h(W_i - w)\,,$$
$$s_{n,j} = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{W_i - w}{h}\right)^j K_h(W_i - w)\,.$$

We write (26) in a matrix form by using a lexicographical order. Let $N_i = \binom{i+d}{d}$ be the number of distinct $(d+1)$-tuples with $|j| = i$, which is the number of distinct derivatives with order i. Order these N_i $(d+1)$-tuples as a sequence in the lexicographical order with the highest priority to the first position, so the sequence starts from $(i, \ldots, 0, 0)$ and ends at $(0, 0, \ldots, i)$. Let g^{-1} denote this one-to-one map. Define

$$\tau_n = \left[\tau_{n,0}^{\intercal}, \tau_{n,1}^{\intercal}, \ldots, \tau_{n,p}^{\intercal}\right]^{\intercal}$$

where $\tau_{n,|j|}$ is a $N_{|j|} \times 1$ vector with $(\tau_{n,|j|})_k = t_{n,g|j|}(k)$. Note that τ_n is of dimension $\mathbf{N} \times 1$ with

$$\mathbf{N} = \sum_{i=0}^{p} N_i\,. \tag{27}$$

Similarly we arrange $h^{|k|}\hat{b}_k$ in the same order to get

$$\hat{\beta}_n = \left[\hat{\beta}_{n,0}^{\intercal} \quad \hat{\beta}_{n,1}^{\intercal} \quad \cdots \quad \hat{\beta}_{n,p}^{\intercal}\right]^{\intercal}.$$

We also arrange the possible values of $s_{n,j+k}$ by a matrix $S_{n,|j|,|k|}$ in a lexicographical order with the (l, m) element $[S_{n,|j|,|k|}]_{l,m} = s_{n,g_{|j|}(l)+g_{|k|}(m)}$. Define the $\mathbf{N} \times \mathbf{N}$ matrix S_n

$$S_n = \begin{bmatrix} S_{n,0,0} & S_{n,0,1} & \cdots & S_{n,0,p} \\ S_{n,1,0} & S_{n,1,1} & \cdots & S_{n,1,p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{n,p,0} & S_{n,0,1} & \cdots & S_{n,p,p} \end{bmatrix}.$$

Before we proceed, we need to introduce some notations following Masry (1996a,b) and Li et al. (2003).

For each j with $0 \le |j| \le d+1$, define

$$\mu_j = \int \upsilon^j K(\upsilon)\, d\upsilon\,, \quad \nu_{s,j} = \int \upsilon_s \upsilon^j K(\upsilon)\, d\upsilon, \quad s = 1, \ldots, d+1\,,$$
$$\kappa_{r,j} = \int \upsilon^r \upsilon^j K(\upsilon)\, d\upsilon, \quad |r| = 2\,,$$

where $\upsilon^j = (\upsilon_1^{j_1}, \cdots, \upsilon_{d+1}^{j_{d+1}})$, $\upsilon_s \upsilon^j = (\upsilon_1^{j_1}, \cdots, \upsilon_s^{1+j_s}, \cdots, \upsilon_{d+1}^{j_{d+1}})$ and $\upsilon^r \upsilon^j = (\upsilon_1^{r_1+j_1}, \cdots, \upsilon_{d+1}^{r_{d+1}+j_{d+1}})$ as defined earlier. Then define $\mathbf{N} \times \mathbf{N}$ matrices M, U_s $(s = 1, \ldots, d+1)$, H_r $(|r| = 2)$, $U(w)$ and $H(w)$ by

$$M = \begin{bmatrix} M_{0,0} & M_{0,1} & \cdots & M_{0,p} \\ M_{1,0} & M_{1,1} & \cdots & M_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ M_{p,0} & M_{p,1} & \cdots & M_{p,p} \end{bmatrix}, U_s = \begin{bmatrix} U_{s,0,0} & U_{s,0,1} & \cdots & U_{s,0,p} \\ U_{s,1,0} & U_{s,1,1} & \cdots & U_{s,1,p} \\ \vdots & \vdots & \ddots & \vdots \\ U_{s,p,0} & U_{s,p,1} & \cdots & U_{s,p,p} \end{bmatrix}, H_r = \begin{bmatrix} H_{r,0,0} & H_{r,0,1} & \cdots & H_{r,0,p} \\ H_{r,1,0} & H_{r,1,1} & \cdots & H_{r,1,p} \\ \vdots & \vdots & \ddots & \vdots \\ H_{r,p,0} & H_{r,p,1} & \cdots & H_{r,p,p} \end{bmatrix} \tag{28}$$

and

$$U(w) = \sum_{s=1}^{d+1} f_s^{(1)}(w) U_s\,, \quad H(w) = \sum_{|r|=2} f_r^{(2)}(w) H_r\,,$$

where $M_{i,j}$, $U_{s,i,j}$ and $H_{r,i,j}$ are $N_i \times N_j$ dimensional matrices whose (l, m) element are $\mu_{g_i(l)+g_j(m)}$, $\nu_{s,g_i(l)+g_j(m)}$ and $\kappa_{r,g_i(l)+g_j(m)}$ respectively, $f_s^{(1)}(w)$ is the sth component of the gradient $f^{(1)}(w)$, and $f_r^{(2)}(w)$ is the second order partial derivative with respect to w^r. Define the centered $t_{n,j}(w)$ as

$$\begin{aligned} t_{n,j}^*(w) &= \frac{1}{n} \sum_{i=1}^n [Y_i - m(W_i)] \left(\frac{W_i - w}{h} \right)^j K_h(W_i - w) \\ &= \frac{1}{n} \sum_{i=1}^n \xi_i \left(\frac{W_i - w}{h} \right)^j K_h(W_i - w)\,. \end{aligned}$$

Then the $\mathbf{N} \times 1$ matrix $\tau_n^*(w)$ are defined the same way as $\tau_n(w)$ but with $t_{n,j}(w)$ replaced with $t_{n,j}^*(w)$.

Use the same lexicographical order as before, we define a column vector $m^{(p+1)}(w)$ as the N_{p+1} elements of derivatives $1/j!(D^j m)(w)$ with $|j| = p+1$. Define the $\mathbf{N} \times N_{p+1}$ matrices $B_n(w)$ and B by

$$B_n(w) = \begin{bmatrix} S_{n,0,p+1} \\ S_{n,1,p+1} \\ \vdots \\ S_{n,p,p+1} \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} M_{0,p+1} \\ M_{1,p+1} \\ \vdots \\ M_{p,p+1} \end{bmatrix}. \tag{29}$$

Similarly as (A.9) in Li et al. (2003), applying Theorem 4 in Masry (1996b), we have (24) becomes

$$
\begin{aligned}
& e_t^{\intercal} \frac{1}{h^2} \mathbb{E}[\hat{\beta}_n(W) - \beta(W)] \\
&= e_t^{\intercal} \frac{1}{h^2} \left[\int S_n^{-1}(w) \tau_n^*(w) \, dF(w) + h^{p+1} \int S_n^{-1}(w) B_n(w) m^{(p+1)}(w) \, dF(w) \right] + O_P(\sqrt{\frac{\log n}{nh^{d+1}}} h^{p-1} + h^p) \,,
\end{aligned} \tag{30}
$$

where $t = 1 + N_1 + d + 1 = 2d + 3$ under the lexicographical order given. For any k^{th} order derivative of the expectation, we shall have $t \in \{1 + N_1 + \cdots + N_{k-1} + 1, \ldots, 1 + N_1 + \cdots + N_k\}$. By Assumption 6(ii), $O_P(\sqrt{\frac{\log n}{nh^{d+1}}} h^{p-1} + h^p) = o_P(n^{-1/2})$.

Masry (1996a,b) has shown that $\sup_{x \in \mathcal{D}} |[S_n(w)]^{-1} - [f(w)M]^{-1}| = o(1)$, and Li et al. (2003) have demonstrated that $\sup_{x \in \mathcal{D}} |[S_n(w)]^{-1} - \{[f(w)M]^{-1} - hG(w)\}| = o(h)$ *a.s.*, where $G(w) = M^{-1}U(w)M^{-1}/f^2(w)$. We prove that higher order result also holds.

Lemma A.2: Under the assumptions of Proposition 1, we have

$$
\sup_{w \in \mathcal{D}} |S_n(w) - f(w)M - hU(w) - h^2 H(w)| = O\left(h^3 + \left(\frac{\log n}{nh^{d+1}} \right)^{1/2} \right) = o(h^2) \,.
$$

Also, the inverse holds that

$$
\sup_{w \in \mathcal{D}} |S_n^{-1}(w) - \{[f(w)M]^{-1} - hG(w) + h^2 Q(w)\}| = o(h^2) \,,
$$

where $Q(w) = [f(w)M]^{-1}U(w)[f(w)M]^{-1}U(w)[f(w)M]^{-1} - [f(w)M]^{-1}H(w)[f(w)M]^{-1}$.

Proof of Lemma A.2: It suffices to show that for each j with $0 \leq |j| \leq \mathbf{N}$, $\sup_{w \in \mathcal{D}} |s_{n,j}(w) - f(w)\mu_j - h \sum_{s=1}^{d+1} f_s^{(1)}(w)\nu_{s,j} - h^2 \sum_{|r|=2} f_r^{(2)}(w)\kappa_{r,j}| = O(h^3 + [\log n/(nh^{d+1})]^{1/2})$. Note that $s_{n,j}(w) = n^{-1} \sum_{i=1}^n ((W_i - w)/h)^j K_h(W_i - w)$, we have

$$
\begin{aligned}
\mathbb{E}[s_{n,j}(w)] &= \int (w_i - w)^j / h^j K_h(w_i - w) f(w_i) \, dw_i \\
&= \int \upsilon^j K(\upsilon) f(w + h\upsilon) \, d\upsilon + f(w) \int \upsilon^j K(\upsilon) \, d\upsilon + h \sum_{s=1}^{d+1} f_s^{(1)}(w) \int \upsilon_s \upsilon^j K(\upsilon) \, d\upsilon \\
&\quad + h^2 \sum_{|r|=2} f_r^{(2)}(w) \int \upsilon_r \upsilon^j K(\upsilon) \, d\upsilon + O(h^3) \,,
\end{aligned}
$$

uniformly in $w \in \mathcal{D} = \mathcal{X} \times \mathcal{V}$. The first result follows from this together with Theorem 2 in Masry

(1996b). Thus, we have

$$S_n(w) = f(w)M + hU(w) + h^2H(w) + O\left(h^3 + \left(\frac{\log n}{nh^{d+1}}\right)^{1/2}\right)$$
$$= f(w)M\{I_n + h[f(w)M]^{-1}U(w) + h^2[f(w)M]^{-1}H(w) + o(h^2)\} \quad a.s.\,,$$

uniformly in $\mathcal{D}$. The second equality is by Assumption 6(ii). It's easy to see that

$$\{I_n + h[f(w)M]^{-1}U(w) + h^2[f(w)M]^{-1}H(w) + o(h^2)\}^{-1}$$
$$= I_n - h[f(w)M]^{-1}U(w) + h^2\{[f(w)M]^{-1}U(w)[f(w)M]^{-1}U(w) - [f(w)M]^{-1}H(w)\} + o(h^2)\,,$$

uniformly in $\mathcal{D}$ by Taylor expansion, as the eigenvalues of the matrix

$$h[f(w)M]^{-1}U(w) + h^2[f(w)M]^{-1}H(w) + o(h^2)$$

are of small order. Then we have

$$S_n^{-1}(w) = [f(w)M]^{-1} - hG(w) + h^2Q(w) + o(h^2)\,.$$

where

$$G(w) = [f(w)M]^{-1}U(w)[f(w)M]^{-1}$$
$$Q(w) = [f(w)M]^{-1}U(w)[f(w)M]^{-1}U(w)[f(w)M]^{-1} - [f(w)M]^{-1}H(w)[f(w)M]^{-1}\,. \blacksquare$$

Applying Lemma A.2 on (30), we shall show that the leading term is indeed determined by higher order terms, as those from the first and second terms of Taylor expansion are degenerate. We have

$$e_t^\intercal \frac{1}{h^2}\mathbb{E}[\hat{\beta}_n(W) - \beta(W)] = e_t^\intercal \frac{1}{h^2}\left\{\int [f(w)M]^{-1}\tau_n^*(w)\,dF(w) - \int hG(w)\tau_n^*(w)\,dF(w) + \int h^2Q(w)\tau_n^*(w)\,dF(w)\right.$$
$$\left. + h^{p+1}\int S_n^{-1}(w)B_n(w)m^{(p+1)}(w)dF(w)\right\} + o_P(n^{-1/2})$$
$$= A_1 - A_2 + A_3 + A_4 + o_P(n^{-1/2})\,.$$

We shall show that A_2 and A_3 contribute to the variance, A_4 is the bias term and the rest terms are

negligible. Similar to the proof of Lemma A.3 in Li et al. (2003)

$$
\begin{aligned}
A_1 &= \int f^{-1}(w)\frac{1}{nh^2}\sum_{i=1}^{n}\sum_{0\le|j|\le p}[M^{-1}]_{t,j}\xi_i\left(\frac{W_i-w}{h}\right)^j K_h(W_i-w)\,dF(w)\\
&= \frac{1}{nh^2}\sum_{i=1}^{n}\xi_i\sum_{0\le|j|\le p}[M^{-1}]_{t,j}\int u^j K(u)\,du\\
&= \frac{1}{nh^2}\sum_{i=1}^{n}\xi_i\sum_{0\le|j|\le p}[M^{-1}]_{t,j}[M]_{j,1} = \frac{1}{nh^2}\sum_{i=1}^{n}\xi_i[M^{-1}M]_{t,1} = 0\,.
\end{aligned}
$$

We shall show A_2 is $O_P(n^{-1/2})$,

$$
\begin{aligned}
A_2 &= \int \frac{1}{nh}\sum_{i=1}^{n}\sum_{0\le|j|\le p}[G(w)]_{t,j}\xi_i\left(\frac{W_i-w}{h}\right)^j K_h(W_i-w)\,dF(w)\\
&= \frac{1}{nh}\sum_{i=1}^{n}\xi_i\int\sum_{0\le|j|\le p}[G(w)]_{t,j}\left(\frac{W_i-w}{h}\right)^j K_h(W_i-w)\,dF(w)\\
&= \frac{1}{nh}\sum_{i=1}^{n}\xi_i\sum_{0\le|j|\le p}\int f^{-1}(W_i-uh)[M^{-1}U(W_i-uh)M^{-1}]_{t,j}u^j K(u)\,du\\
&= \frac{1}{nh}\sum_{i=1}^{n}\xi_i f^{-1}(W_i)\sum_l[M^{-1}U(W_i)]_{t,l}\sum_{0\le|j|\le p}[M^{-1}]_{l,j}\int u^j K(u)\,du\\
&\quad -\frac{1}{n}\sum_{i=1}^{n}\xi_i\int\sum_{s'=1}^{d+1}(f^{-1})^{(1)}_{s'}(W_i)\sum_{0\le|j|\le p}[M^{-1}U(W_i)M^{-1}]_{t,j}u_{s'}u^j K(u)\,du\\
&\quad -\frac{1}{n}\sum_{i=1}^{n}\xi_i\sum_{0\le|j|\le p}\int f^{-1}(W_i)\sum_{s'=1}^{d+1}[M^{-1}(U(W_i))^{(1)}_{s'}M^{-1}]_{t,j}u_{s'}u^j K(u)\,du + o_P(n^{-1/2})\,,
\end{aligned}
$$

where we have used Taylor expansion for $f^{-1}(W_i-uh)$ and $U(W_i-uh)$. By the definition of M and

$U_s(W_i)$, we have

$$
\begin{aligned}
A_2 &= \frac{1}{nh}\sum_{i=1}^{n}\xi_i f^{-1}(W_i)\sum_{k}[M^{-1}]_{t,k}\sum_{s=1}^{d+1}f_s^{(1)}(W_i)\sum_{l}[U_s]_{k,l}I_{l,1} \\
&\quad - \frac{1}{n}\sum_{i=1}^{n}\xi_i\int\sum_{s'=1}^{d+1}\sum_{s=1}^{d+1}(f^{-1})_{s'}^{(1)}(W_i)f_s^{(1)}(W_i)\sum_{0\le|j|\le p}[M^{-1}U_sM^{-1}]_{t,j}u_{s'}u^j K(u)\,du \\
&\quad - \frac{1}{n}\sum_{i=1}^{n}\xi_i\sum_{0\le|j|\le p}\int f^{-1}(W_i)\sum_{s=1}^{d+1}\sum_{s'=1}^{d+1}f_{ss'}^{(2)}(W_i)[M^{-1}U_sM^{-1}]_{t,j}u_{s'}u^j K(u)\,du + o_P(n^{-1/2}) \\
&= \frac{1}{nh}\sum_{i=1}^{n}\xi_i f^{-1}(W_i)\sum_{s=1}^{d+1}f_s^{(1)}(W_i)\sum_{k}[M^{-1}]_{t,k}[U_s]_{k,1} \\
&\quad - \frac{1}{n}\sum_{i=1}^{n}\xi_i\sum_{s=1}^{d+1}\sum_{s'=1}^{d+1}[(f^{-1})_{s'}^{(1)}(W_i)f_s^{(1)}(W_i)+f^{-1}(W_i)f_{ss'}^{(2)}(W_i)]\sum_{0\le|j|\le p}[M^{-1}U_sM^{-1}]_{t,j}[U_{s'}]_{j,1} + o_P(n^{-1/2}) \\
&= A_{21} - A_{22} + o_P(n^{-1/2})\,.
\end{aligned}
$$

The first term A_{21} is degenerate, as for any k, we can show $[M^{-1}]_{t,k}[U_s]_{k,1}=0$. It's sufficient to show $[M^{-1}]_{t,k}=0$, whenever $[U_s]_{k,1}\neq 0$. We have that for $s=1,...,d+1$,

$$
\begin{aligned}
U_s[\cdot,1] &= [0\ \square_{1\times N_1}\ 0_{1\times N_2}\ \square_{1\times N_3}\ 0_{1\times N_4}\ \dots\ 0_{1\times N_p}]^{\intercal}\,, \quad \text{if p is even}\,, \\
U_s[\cdot,1] &= [0\ \square_{1\times N_1}\ 0_{1\times N_2}\ \square_{1\times N_3}\ 0_{1\times N_4}\ \dots\ \square_{1\times N_p}]^{\intercal}\,, \quad \text{if p is odd}\,,
\end{aligned}
$$

where $\square_{1\times l}$ represents a $1\times l$ nonzero row vector. Note that $N_0=1$. As we use product kernel, it's easy to see that when p is odd, M is a block matrix with zero blocks and nonzero blocks appear alternatively, i.e.

$$
M=\begin{bmatrix}
\square_{N_0\times N_0} & 0_{N_0\times N_1} & \square_{N_0\times N_2} & \dots & 0_{N_0\times N_p} \\
0_{N_1\times N_0} & \square_{N_1\times N_1} & 0_{N_1\times N_2} & \dots & \square_{N_1\times N_p} \\
\vdots & \vdots & \vdots & \dots & \vdots \\
0_{N_p\times N_0} & \square_{N_p\times N_1} & 0_{N_p\times N_2} & \dots & 0_{N_p\times N_p}
\end{bmatrix}.
$$

When p is even, it is clear to see M has the same pattern. We use the result of the lemma below.

Lemma A.3: M^{-1} has the same zero blocks as M.

Proof of Lemma A.3: We only prove the result when p is odd. When p is even, the derivation is similar. Let $p=2q-1$, $N_e=N_0+N_2+\cdots+N_{p-1}$ and $N_e=N_1+N_3+\cdots+N_p$. We first apply the

row switching block elementary matrix $T_{i,j}$ on M, then we have

$$T_{N_1,N_{p-1}}T_{N_3,N_{p-3}}\cdots T_{N_{q-1},N_q}MT_{N_{q-1},N_q}\cdots T_{N_3,N_{p-3}}T_{N_1,N_{p-1}} = \begin{bmatrix} A_{N_e\times N_e} & 0_{N_e\times N_o} \\ 0_{N_o\times N_e} & B_{N_o\times N_o} \end{bmatrix},$$

where A and B are invertible matrices. Then, we have

$$M^{-1} = (T_{N_1,N_{p-1}}T_{N_3,N_{p-3}}\cdots T_{N_{q-1},N_q})^{-1}\begin{bmatrix} A^{-1}_{N_e\times N_e} & 0_{N_e\times N_o} \\ 0_{N_o\times N'_e} & B^{-1}_{N_o\times N_o} \end{bmatrix}(T_{N_{q-1},N_q}\cdots T_{N_3,N_{p-3}}T_{N_1,N_{p-1}})^{-1}. \blacksquare$$

For the tth row of M, it belongs to the N_0+N_1+1 to $N_0+N_1+N_2$ block, so we have

$$[M]_{t,\cdot} = \begin{bmatrix} \square_{1\times N_0} & 0_{1\times N_1} & \square_{1\times N_2}\ldots \end{bmatrix}.$$

Thus, it's obvious that A_{21} is degenerate. For A_{22}, we have

$$\begin{aligned} A_{22} &= \frac{1}{n}\sum_{i=1}^{n}\xi_i\sum_{s=1}^{d+1}\sum_{s'=1}^{d+1}[(f^{-1})^{(1)}_{s'}(W_i)f^{(1)}_s(W_i) + f^{-1}(W_i)f^{(2)}_{ss'}(W_i)][M^{-1}U_sM^{-1}U_{s'}]_{t,1} \\ &= \frac{1}{n}\sum_{i=1}^{n}\xi_i[R_1(W_i)]_{t,1}, \end{aligned}$$

where $R_1(w) = \sum_{s=1}^{d+1}\sum_{s'=1}^{d+1}[(f^{-1})^{(1)}_{s'}(w)f^{(1)}_s(w) + f^{-1}(w)f^{(2)}_{ss'}(w)]M^{-1}U_sM^{-1}U_{s'}$.

Finally, we have

$$\begin{aligned} A_3 &= \frac{1}{n}\sum_{i=1}^{n}\xi_i\sum_{0\le|j|\le p}\int[Q(w)]_{t,j}\left(\frac{W_i-w}{h}\right)^j K_h(W_i-w)\,dF(w) \\ &= \frac{1}{n}\sum_{i=1}^{n}\xi_i\sum_{0\le|j|\le p}\int[Q(W_i-uh)]_{t,j}f(W_i-uh)u^jK(u)\,du \\ &= \frac{1}{n}\sum_{i=1}^{n}\xi_i\sum_{0\le|j|\le p}[Q(W_i)]_{t,j}\mu_jf(W_i) + o_P(n^{-1/2}) = \frac{1}{n}\sum_{i=1}^{n}\xi_i[R_2(W_i)]_{t,1} + o_P(n^{-1/2}), \end{aligned}$$

where $R_2(w) = Q(w)f(w)M$. Combining A_{22} and A_3, let

$$R(w) = R_1(w) + R_2(w) = \sum_{s=1}^{d+1}\sum_{s'=1}^{d+1}[(f^{-1})^{(1)}_{s'}(w)f^{(1)}_s(w) + f^{-1}(w)f^{(2)}_{ss'}(w)]M^{-1}U_sM^{-1}U_{s'} + Q(w)f(w)M. \tag{31}$$

Finally, applying the uniform results in Masry (1996b), we have

$$\begin{aligned} A_4 &= e_t^\intercal h^{p-1} \int S_n^{-1}(w) B_n(w) m^{(p+1)}(w)\, dF(w) \\ &= e_t^\intercal h^{p-1} M^{-1} B \mathbb{E}[m^{(p+1)}(w)] + O_P(h^p)\,. \end{aligned}$$

Altogether we have

$$\sqrt{n}\mathbb{E}\left[\frac{\widehat{\partial^2 \mathbb{E}}(Y|X,V)}{\partial X_1 \partial V} - \frac{\partial^2 \mathbb{E}(Y|X,V)}{\partial X_1 \partial V} - e_t^\intercal h^{p-1} M^{-1} B \mathbb{E}[m^{(p+1)}(w)]\right] = \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i [R(W_i)]_{t,1}\,,$$

then we can apply CLT directly for the righthand side.

Now, let's check for the order of s.o.$(\hat{f})$. From Masry (1996a,b), we have

$$\|\hat{\pi} - \pi\|_\infty = O_P\left([\log n/(n g^{d_z})]^{1/2} + g^2\right)\,.$$

Therefore, $O(\|\hat{f}_2 - \bar{f}_2\|_\infty^2) = o_P(n^{-1/2})$ by Assumption 6(i). By a similar argument as in Masry (1996a,b) or Mammen et al. (2012a), we can show that we have $\|\hat{f}_2 - \bar{f}_2\|_\infty \|\hat{f}_1^{(v)} - \bar{f}_1^{(v)}\|_\infty = O_P(\log n/(n^2 g^{dz} h^{d+7})^{1/2}) = o_P(n^{-1/2})$ by Assumption 6(iii). Therefore, s.o.$(\hat{f}) = o_P(n^{-1/2})$.

To show $T_{2,n}(\hat{f}) = O_P(n^{-1/2})$, recall that

$$T_{2,n}(\hat{f}) = \frac{1}{n} \sum_{i=1}^n \frac{\partial^3 \mathbb{E}(Y_i|X_i,V_i)}{\partial X_1 \partial V^2} (\hat{V}_i - V_i)\,.$$

By Lemma A.4 below, we can show that

$$T_{2,n}(\hat{f}) = \mathbb{E}\left[(\hat{V} - V) \frac{\partial^3 \mathbb{E}(Y|X,V)}{\partial X_1 \partial V^2}\right] + o_P(n^{-1/2})\,.$$

Lemma A.4: Under the assumptions of Proposition 1, we have

$$\sup_{V_1, V_2 \in \bar{\mathcal{M}}_n} \left| \frac{1}{n} \sum_{i=1}^n [V_1(X_i, Z_i) - V_2(X_i, Z_i)] \frac{\partial^3 \mathbb{E}(Y_i|X_i,V_i)}{\partial X_1 \partial V^2} - \mathbb{E}\left[(V_1 - V_2) \frac{\partial^3 \mathbb{E}(Y|X,V)}{\partial X_1 \partial V^2}\right] \right| = o_P(n^{-1/2})\,,$$

Proof of Lemma A.4: Define $\Delta_i(V_1, V_2) = [V_1(X_i, Z_i) - V_2(X_i, Z_i)] \frac{\partial^3 \mathbb{E}(Y_i|X_i,V_i)}{\partial X \partial V^2} - \mathbb{E}[(V_1 - V_2) \frac{\partial^3 \mathbb{E}(Y|X,V)}{\partial X \partial V^2}]$. Then the proof follows closely of Lemma A.5 in Lee (2013), which modifies that of Lemma 1 in Mammen et al. (2012a). We need to verify that Assumption 2 (Accuracy) and Assumption 3 (Complexity) in Mammen et al. (2012a) hold, which are

- (Accuracy) $\sup_z |\hat{\pi}(z) - \pi(z)| = o_P(n^{-\iota})$ for some ι such that $n^{-\iota} = o(h)$.
- (Complexity) There exists sequences of sets $\mathcal{M}_n$, such that
  1. $\hat{\pi}(\cdot) \in \mathcal{M}_n$.
  2. For a constant $C_M > 0$ and a function π_n with $\|\pi_n - \pi\|_\infty = o(n^{-\iota})$, the set $\bar{\mathcal{M}}_n = \mathcal{M}_n \cap \{\pi^\star : \|\pi^\star - \pi_n\|_\infty \leq n^{-\iota}\}$ can be covered by at most $C_M \exp(c_\lambda^{-\psi} n^\zeta)$ balls with $\|\cdot\|_\infty$-radius c_λ for all $c_\lambda \leq n^{-\iota}$, where $0 < \psi \leq 2$, $\zeta \in \mathcal{R}$.

and also $\iota - \frac{1}{2}(\iota\psi + \zeta) > 0$, which indicates $\kappa_1 > 1/2$ in Lemma A.3 in Lee (2013). As we use local linear regression for $\hat{\pi}(z)$, we know $n^{-\iota} = o(h)$ by Assumption 6 and we can let $\mathcal{M}_n$ be the set of functions defined on the compact support of Z with up to α order partial derivatives and uniformly bounded by some multiple of $n^{\zeta^\star}$ with $\zeta^\star \geq 0$. Then, by Corollary 2.7.2 in van der Vaart and Wellner (1996), we have $\psi = \frac{d_z}{\alpha}$ and $\zeta = \zeta^\star \psi$. By choosing $\zeta^\star$ sufficiently small and $d_z < 2\alpha$, the result holds. ■

We know that $\hat{V} - V = -[\hat{\pi}(Z) - \pi(Z)]$, then

$$\sqrt{n} T_{2,n}(\hat{f}) = -\frac{1}{\sqrt{n}} \sum_{i=1}^n V_i \mathbb{E}_{X|Z_i} \left\{ \frac{\partial^3 \mathbb{E}[Y|X, X - \pi(Z_i)]}{\partial X_1 \partial V^2} \right\} + o_P(1) .$$

Altogether, we have by (21),

$$\begin{aligned} \sqrt{n}\{S_n(\hat{f}) - 2e_t' h^{p-1} M^{-1} B \mathbb{E}[m^{(p+1)}(w)]\} = \frac{2}{\sqrt{n}} \sum_{i=1}^n \xi_i [R(W_i)]_{t,1} - \frac{1}{\sqrt{n}} \sum_{i=1}^n V_i \mathbb{E}_{X|Z_i} \left\{ \frac{\partial^3 \mathbb{E}[Y|X, X - \pi(Z_i)]}{\partial X_1 \partial V^2} \right\} \\ + \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \frac{\partial^2 \mathbb{E}(Y_i|X_i, V_i)}{\partial X_1 \partial V} - \mathbb{E}\left[\frac{\partial^2 \mathbb{E}(Y|X,V)}{\partial X_1 \partial V} \right] \right\} + o_P(1) . \end{aligned}$$

Thus, the desired result follows by the CLT. ■

Let $m^D(w) = \mathbb{E}(D|W = w)$, the following can be proved in the same way as Proposition 1.

Corollary A.1: Under assumptions of Theorem 1, we have

$$\begin{aligned} \sqrt{n} \left\{ \hat{\mathbb{E}} \left[\frac{\widehat{\partial^2 \mathbb{E}}(D|X, \hat{V})}{\partial X_1 \partial V} \right] - \mathbb{E}\left[\frac{\partial^2 \mathbb{E}(D|X,V)}{\partial X_1 \partial V} \right] - 2e_t' h^{p-1} M^{-1} B \mathbb{E}[m_{p+1}^D(X,V)] \right\} = \frac{2}{\sqrt{n}} \sum_{i=1}^n u_i [R(X_i, V_i)]_{t,1} \\ - \frac{1}{\sqrt{n}} \sum_{i=1}^n V_i \mathbb{E}_{X|Z_i} \left\{ \frac{\partial^3 \mathbb{E}[D|X, X - \pi(Z_i)]}{\partial X_1 \partial V^2} \right\} + \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ \frac{\partial^2 \mathbb{E}(D_i|X_i, V_i)}{\partial X_1 \partial V} - \mathbb{E}\left[\frac{\partial^2 \mathbb{E}(D|X,V)}{\partial X_1 \partial V} \right] \right\} + o_P(1) . \end{aligned}$$

Proof of Theorem 1: By Delta method and the results of Proposition 1 and Corollary A.1, we have

$$\sqrt{n}(\hat{\gamma}-\gamma)=\frac{1}{\mathbb{E}\left[\frac{\partial^2\mathbb{E}(D|X,V)}{\partial X_1\partial V}\right]}\frac{2}{\sqrt{n}}\sum_{i=1}^{n}\varsigma_i[R(X_i,V_i)]_{t,1}+o_P(1)\,.$$

Thus by CLT, we have the desired result. ■

Proof of Corollary 1: By the proof of Proposition 1 and the fact

$$\frac{\partial^2\mathbb{E}(Y|X,V)}{\partial X_1\partial V}=\gamma\left[\frac{\partial^2\mathbb{E}(D|X,V)}{\partial X_1\partial V}\right]\quad\text{and}\quad\frac{\partial^3\mathbb{E}(Y|X,V)}{\partial X_1\partial V^2}=\gamma\left[\frac{\partial^3\mathbb{E}(D|X,V)}{\partial X_1\partial V^2}\right],$$

it's easy to see that any additional term caused by generated regressor is cancelled. Therefore, we have the identical result as Theorem 1. ■